

APPLICATION
FOR
UNITED STATES LETTERS PATENT

TITLE: PLANT RETROELEMENTS AND METHODS RELATED
THERETO

APPLICANT: DAVID A. WRIGHT AND DANIEL F. VOYTAS

CERTIFICATE OF MAILING BY EXPRESS MAIL

Express Mail Label No. EV 342626258 US

July 8, 2003

Date of Deposit

PLANT RETROELEMENTS AND METHODS RELATED THERETO

This application claims priority to US Provisional Patent Application Serial Number 60/087125, filed May 29, 1998.

10 The present invention was funded, in part, by the United States Department of Agriculture, Contract Number IOW03120; the United States Government may have certain rights in the invention.

FIELD OF THE INVENTION

15

The present invention provides plant retroelements and methods related to plant retroelements. The invention involves techniques from the fields of: molecular biology, virology, genetics, bioinformatics, and, to a lesser extent, other related fields.

20

BACKGROUND OF THE INVENTION

25

The eukaryotic retrotransposons are divided into two distinct classes of elements based on their structure: the long terminal repeat (LTR) retrotransposons and the LINE-like or non LTR elements. Doolittle et al. (1989) Quart. Rev. Biol. 64: 1-30; Xiong and Eickbush (1990) EMBO J 9: 3353-3362. These element classes are related by the fact that each must undergo reverse transcription of an RNA intermediate to replicate, and each generally encodes its own reverse transcriptase. The LTR retrotransposons replicate by a mechanism which resembles that of the retroviruses. Boeke and Sandmeyer, (1991) Yeast transposable elements. In The Molecular and Cellular Biology of the Yeast *Saccharomyces*, edited by J. Broach, E. Jones and J. Pringle, pp. 193-261. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y. They typically use a specific tRNA to prime reverse transcription, and a linear cDNA is synthesized through a series of template transfers that require redundant LTR sequences at each end of the element mRNA. This all occurs within a virus-like particle formed from proteins encoded by the retrotransposon mRNA. After reverse transcription, an integration complex is organized that directs the resulting cDNA to a new site in the genome of the host cell.

30

5 Phylogenetic analyses based on reverse transcriptase amino acid sequences resolve the LTR retrotransposons into two families: the Ty3/gypsy retrotransposons (Metaviridae), and the Ty1/copia elements (Pseudoviridae).
Boeke et al., (1998) Metaviridae. In Virus Taxonomy: ICTV VIIth Report, edited by F. A. Murphy. Springer-Verlag, New York; Boeke et al. (1998) Pseudoviridae.
In Virus Taxonomy: ICTV VIIth Report, edited by F. A. Murphy. Springer
Verlag, New York.; Xiong and Eickbush (1990) EMBO J. 9: 3353-3362.
10 Although distinct, Ty3/gypsy elements are more closely related to the retroviruses than to the Ty1/copia elements. They also share a similar genetic organization with the retroviruses, principally in the order of integrase and reverse transcriptase in their pol genes. For the Ty3/gypsy elements, reverse transcriptase precedes integrase, and this order is reversed for the Ty1/copia elements. In addition, some Ty3/gypsy elements have an extra open reading frame (ORF) similar to retroviral
15 envelope (env) proteins, which is required for viral infectivity. The *Drosophila melanogaster* gypsy retrotransposons encode an env-like ORF and can be transmitted between cells. Kim et al. (1994) Proc. Natl. Acad. Sci. USA 91: 1285-1289; Song et al. (1994) Genes & Dev. 8: 2046-2057. Thus there are two distinct lineages of infectious LTR retroelements, the retroviruses and those Ty3/gypsy retrotransposons that encode envelope-like proteins. The Ty3/gypsy elements have
20 been divided into two genera, the metaviruses and the errantiviruses, the latter of which include all elements with env-like genes. Boeke et al., (1998) Metaviridae.
In Virus Taxonomy: ICTV VIIth Report, edited by F. A. Murphy. Springer-Verlag, New York

25 In plants, retrotransposons have been extremely successful. Bennetzen (1996) Trends Microbiol. 4: 347-353; Voytas (1996) Genetics 142: 569-578. The enormous size of many plant genomes demonstrates a great tolerance for repetitive DNA, a substantial proportion of which appears to be composed of retrotransposons. Because of their abundance, retrotransposons have undoubtedly influenced plant gene evolution. They can cause mutations in coding sequences (Grandbastien et al. (1989) Nature 337: 376-380; Hirochika et al. (1996) Proc. Natl. Acad. Sci. USA 93: 7783-7788; Purugganan and Wessler (1994) Proc. Natl. Acad. Sci. USA 91: 11674-11678), and the promoter regions of some plant genes contain relics of retrotransposon insertions that contribute transcriptional regulatory sequences. White et al. (1994) Proc. Natl. Acad. Sci. USA 91: 11792-11796. Retrotransposons also generate gene duplications: Repetitive retrotransposon
30
35

5 sequences provide substrates for unequal crossing over, and such an event is thought to have caused a zein gene duplication in maize. White et al. (1994) Proc. Natl. Acad. Sci. USA 91: 11792-11796. Occasionally, cellular mRNAs are reverse transcribed and the resultant cDNA recombines into the genome giving rise
10 to new genes, or more frequently, cDNA pseudogenes. Maestre et al. (1995) EMBO J. 14: 6333-6338. The transduction of gene sequences during reverse transcription, which produced the oncogenic retroviruses, has also been documented to occur for a plant retrotransposon (Bureau et al. (1994) Cell 77: 479-480.; Jin and Bennetzen (1994) Plant Cell 6: 1177 1186); a maize Bs1 insertion in Adh1 carries part of an ATPase gene and is the only known example of a retrotransposon-mediated gene transduction event.

15 Plant genomes encode representatives of the two major lineages of LTR retrotransposons that have been identified in other eukaryotes. Among these are numerous examples of Ty1/copia elements (e.g. Konieczny et al. (1991) Genetics 127: 801-809; Voytas and Ausubel (1988) Nature 336: 242-244; Voytas et al.
20 (1990) Genetics 126: 713-721) Also prevalent are Ty3/gypsy elements, which are members of the genus Metaviridae (Smyth et al. 1989; Purugganan and Wessler 1994 Proc. Natl. Acad. Sci. USA 91: 11674-11678; Su and Brown 1997). As stated above, the metaviruses do not encode an envelope protein characteristic of the retroviruses. It has been suggested that some plant retrovirus-like elements may have lost, or not yet gained, genes such as the envelope gene required for cell-to-cell transmission (Bennetzen (1996) Trends Microbiol. 4: 347-353). As one group
25 of researchers described the uncertainty, “[s]ince genes encoding ENV [envelope] functions are very heterogeneous at the sequence level and difficult to identify by homology even between retroviruses, the possibility cannot be completely excluded at the present time that the 3' ORF of Cyclops [the retrotransposon described in the paper] is, in fact, an env gene and, hence, Cyclops is a retrovirus or a descendant of one.” Chavanne et al. (1998) Plant Molecular Biol 37: 363-375.

30 Citation of the above documents is not intended as an admission that any of the foregoing is pertinent prior art. All statements as to the date or representation as to the contents of these documents is based on subjective characterization of information available to the applicant, and does not constitute any admission as to the accuracy of the dates or contents of these documents.

SUMMARY OF THE INVENTION

5 In general, the present invention provides materials, such as nucleic acids, vectors, cells, and plants (including plant parts, seeds, embryos, etc.), and methods to manipulate the materials. In particular, molecular tools are provided in the form of retroelements and retroelement-containing vectors, cells and plants. The particular methods include methods to introduce the retroelements into cells, especially wherein the retroelements carries at least one agronomically-significant characteristic. The best mode of the present invention is a particular method to transfer agronomically-significant characteristics to plants wherein a helper cell line 10 which expresses gag, pol and env sequences is used to enable transfer of a secondary construct which carries an agronomically-significant characteristic and has retroelement sequences that allow for replication and integration.

15 In one embodiment, there are provided isolated nucleic acid molecules, wherein said nucleic acid molecules encode at least a portion of a plant retroelement and comprises a nucleic acid sequence selected from the group consisting of:

20 (a) a nucleic acid sequence which is a plant retroelement primer binding site and which has more than 95% identity to SEQ ID NO 2, wherein said identity can be determined using the DNAsis computer program and default parameters;

25 (b) a nucleic acid sequence which is at least a portion of a plant retroelement envelope sequence and which has more than 50% identity to SEQ ID NO 5, wherein said identity can be determined using the DNAsis computer program and default parameters;

30 (c) a nucleic acid sequence which is at least a portion of a plant retroelement gag sequence and which has more than 50% identity to SEQ ID NO 7, wherein said identity can be determined using the DNAsis computer program and default parameters;

35 (d) a nucleic acid sequence which is at least a portion of a plant retroelement integrase sequence and which has more than 70% identity to SEQ ID NO 9, wherein said identity can be determined using the DNAsis computer program and default parameters;

5 (e) a nucleic acid sequence which is at least a portion of a plant retroelement reverse transcriptase sequence and which has more than 70% identity to SEQ ID NO 11, wherein said identity can be determined using the DNAsis computer program and default parameters;

10 (f) a nucleic acid sequence which is at least a portion of a plant retroelement protease sequence and which has more than 50% identity to SEQ ID NO 13, wherein said identity can be determined using the DNAsis computer program and default parameters;

15 (g) a nucleic acid sequence which is at least a portion of a plant retroelement RNaseH sequence and which has more than 70% identity to SEQ ID NO 15, wherein said identity can be determined using the DNAsis computer program and default parameters;

20 (h) a nucleic acid sequence which is at least a portion of a plant retroelement sequence and which has more than 50% identity to SEQ ID NO 17, wherein said identity can be determined using the DNAsis computer program and default parameters;

25 (i) a nucleic acid sequence which is selected from the group consisting of: SEQ ID NO 2; SEQ ID NO 5; SEQ ID NO 7; SEQ ID NO 9; SEQ ID NO 11; SEQ ID NO 13; SEQ ID NO 15; and SEQ ID NO 17.

30 (j) a nucleic acid sequence which encodes an amino acid sequence which is at least a portion of a plant retroelement envelope sequence and has more than 30% identity to SEQ ID NO 6, wherein said identity can be determined using the DNAsis computer program and default parameters;

35 (k) a nucleic acid sequence which encodes an amino acid sequence which is at least a portion of a plant retroelement gag sequence and has more than 30% identity to SEQ ID NO 8, wherein said identity can be determined using the DNAsis computer program and default parameters;

(l) a nucleic acid sequence which encodes an amino acid sequence which is at least a portion of a plant retroelement integrase sequence and has more than 75% identity

to SEQ ID NO 10, wherein said identity can be determined using the DNAsis computer program and default parameters;

5 (m) a nucleic acid sequence which encodes an amino acid sequence which is at least a portion of a plant retroelement reverse transcriptase sequence and has more than 79% identity to SEQ ID NO 12, wherein said identity can be determined using the DNAsis computer program and default parameters;

10 (n) a nucleic acid sequence which encodes an amino acid sequence which is at least a portion of a plant retroelement protease sequence and has more than 55% identity to SEQ ID NO 14, wherein said identity can be determined using the DNAsis computer program and default parameters;

15 (o) a nucleic acid sequence which encodes an amino acid sequence which is at least a portion of a plant retroelement RNaseH sequence and has more than 90% identity to SEQ ID NO 16, wherein said identity can be determined using the DNAsis computer program and default parameters;

20 (p) a nucleic acid sequence which encodes an amino acid sequence which is at least a portion of a plant retroelement sequence and has more than 40% identity to SEQ ID NO 18, wherein said identity can be determined using the DNAsis computer program;

25 (q) a nucleic acid sequence which encodes an amino acid sequence selected from the group consisting of: SEQ ID NO 4; SEQ ID NO 6; SEQ ID NO 8; SEQ ID NO 10; SEQ ID NO 12; SEQ ID NO 14; SEQ ID NO 16; and SEQ ID NO 18;

30 (r) a nucleic acid sequence which encodes an allelic variant of an amino acid sequence selected from the group consisting of: SEQ ID NO 4; SEQ ID NO 6; SEQ ID NO 8; SEQ ID NO 10; SEQ ID NO 12; SEQ ID NO 14; SEQ ID NO 16; and SEQ ID NO 18; and

35 (s) a nucleic acid sequence fully complementary to a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); a nucleic acid sequence of (b); a nucleic acid sequence of (c); a nucleic acid sequence of (d); a nucleic acid sequence of (e); a nucleic acid sequence of (f); a nucleic acid sequence of (g); a

5 nucleic acid sequence of (h); a nucleic acid sequence of (i); a nucleic acid sequence of (j); a nucleic acid sequence of (k); a nucleic acid sequence of (l); a nucleic acid sequence of (m); a nucleic acid sequence of (n); a nucleic acid sequence of (o); a nucleic acid sequence of (p); a nucleic acid sequence of (q); and a nucleic acid sequence of (r).

10 Seeds and plants comprising a nucleic acid as above are particularly provided. Nucleic acid molecules as above which comprise gag, pol and env genes and which comprise adenine-thymidine-guanidine as the gag gene start codon are also particularly provided. Those which comprise gag, pol and env genes, the adenine-thymidine-guanidine as the gag gene start codon, and which further comprises SEQ ID NO 4 are also provided.

15 Plant envelope sequences and constructs which comprise the sequences are provided, as are cells, seeds, embryos and plants comprising them. Preferred are isolated nucleic acid molecules, wherein said nucleic acid molecules encode at least a portion of a plant envelope sequence and comprises a nucleic acid sequence selected from the group consisting of:

20 (a) a nucleic acid sequence which has more than 90% identity to SEQ ID NO 5, wherein said identity can be determined using the DNAsis computer program and default parameters;

25 (b) a nucleic acid sequence which encodes SEQ ID NO 5;

(c) a nucleic acid sequence which encodes an amino acid sequence which has greater than 85% identity to SEQ ID NO 6, wherein said identity can be determined using the DNAsis computer program and default parameters;

30 (d) a nucleic acid sequence which encodes amino acid sequence SEQ ID NO 6;

(e) a nucleic acid sequence which encodes an allelic variant of SEQ ID NO 6; and

(f) a nucleic acid sequence fully complementary to a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); a nucleic acid sequence of (b); a nucleic acid sequence of (c); a nucleic acid sequence of (d); and a nucleic acid sequence of (e).

5

Plant cells comprising an isolated nucleic acid molecule above are particularly preferred. Also preferred are plant envelope proteins comprising an amino acid sequence encoded by the above. Methods to impart agronomically-significant characteristics to at least one plant cell are also provided, comprising: contacting a plant envelope protein as described to at least one plant cell under conditions sufficient to allow a nucleic acid molecule to enter said cell, wherein said nucleic acid molecule encodes an agronomically-significant characteristic.

10

Plant integrase sequences and constructs which comprise the sequences are provided, as are cells, seeds, embryos and plants comprising them. Preferred are isolated nucleic acid molecules, wherein said nucleic acid molecules encode at least a portion of a plant integrase sequence and comprises a nucleic acid sequence selected from the group consisting of:

15

(a) a nucleic acid sequence which has more than 90% identity to SEQ ID NO 9, wherein said identity can be determined using the DNAsis computer program and default parameters;

(b) a nucleic acid sequence which encodes SEQ ID NO 9;

20

(c) a nucleic acid sequence which encodes an amino acid sequence which has greater than 85% identity to SEQ ID NO 10, wherein said identity can be determined using the DNAsis computer program and default parameters;

25

(d) a nucleic acid sequence which encodes amino acid sequence SEQ ID NO 10;

(e) a nucleic acid sequence which encodes an allelic variant of SEQ ID NO 10; and

5 (f) a nucleic acid sequence fully complementary to a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); a nucleic acid sequence of (b); a nucleic acid sequence of (c); a nucleic acid sequence of (d); and a nucleic acid sequence of (e).

10 5 Plant cells comprising an isolated nucleic acid molecule above are particularly preferred. Also preferred are plant integrase proteins comprising an amino acid sequence encoded by the above. Methods to impart agronomically-significant characteristics to at least one plant cell are also provided, comprising: contacting a plant integrase protein as described to at least one plant cell under conditions sufficient to allow a nucleic acid molecule to enter said cell, wherein said nucleic acid molecule encodes an agronomically-significant characteristic.

15 15 Plant reverse transcriptase sequences and constructs which comprise the sequences are provided, as are cells, seeds, embryos and plants comprising them. Preferred are isolated nucleic acid molecules, wherein said nucleic acid molecules encode at least a portion of a plant reverse transcriptase sequence and comprises a nucleic acid sequence selected from the group consisting of:

20 20 (a) a nucleic acid sequence which has more than 90% identity to SEQ ID NO 11, wherein said identity can be determined using the DNAsis computer program and default parameters;

25 (b) a nucleic acid sequence which encodes SEQ ID NO 11;

(c) a nucleic acid sequence which encodes an amino acid sequence which has greater than 85% identity to SEQ ID NO 12, wherein said identity can be determined using the DNAsis computer program and default parameters;

30 (d) a nucleic acid sequence which encodes amino acid sequence SEQ ID NO 12;

(e) a nucleic acid sequence which encodes an allelic variant of SEQ ID NO 12; and

5 (f) a nucleic acid sequence fully complementary to a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); a nucleic acid sequence of (b); a nucleic acid sequence of (c); a nucleic acid sequence of (d); and a nucleic acid sequence of (e).

5

10 Plant cells comprising an isolated nucleic acid molecule above are particularly preferred. Also preferred are plant reverse transcriptase proteins comprising an amino acid sequence encoded by the above. Methods to impart agronomically-significant characteristics to at least one plant cell are also provided, comprising: contacting a plant reverse transcriptase protein as described to at least one plant cell under conditions sufficient to allow a nucleic acid molecule to enter said cell, wherein said nucleic acid molecule encodes an agronomically-significant characteristic.

10

15 Plant RNaseH sequences and constructs which comprise the sequences are provided, as are cells, seeds, embryos and plants comprising them. Preferred are isolated nucleic acid molecules, wherein said nucleic acid molecules encode at least a portion of a plant RNaseH sequence and comprises a nucleic acid sequence selected from the group consisting of:

20

(a) a nucleic acid sequence which has more than 90% identity to SEQ ID NO 15, wherein said identity can be determined using the DNAsis computer program and default parameters;

25

(b) a nucleic acid sequence which encodes SEQ ID NO 15;

(c) a nucleic acid sequence which encodes an amino acid sequence which has greater than 95% identity to SEQ ID NO 16, wherein said identity can be determined using the DNAsis computer program and default parameters;

30

(d) a nucleic acid sequence which encodes amino acid sequence SEQ ID NO 16;

35

(e) a nucleic acid sequence which encodes an allelic variant of SEQ ID NO 16; and

5 (f) a nucleic acid sequence fully complementary to a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); a nucleic acid sequence of (b); a nucleic acid sequence of (c); a nucleic acid sequence of (d); and a nucleic acid sequence of (e).

10 5 Plant cells comprising an isolated nucleic acid molecule above are particularly preferred. Also preferred are plant RNaseH proteins comprising an amino acid sequence encoded by the above. Methods to impart agronomically-significant characteristics to at least one plant cell are also provided, comprising: contacting a plant RNaseH protein as described to at least one plant cell under conditions sufficient to allow a nucleic acid molecule to enter said cell, wherein said nucleic acid molecule encodes an agronomically-significant characteristic.

15 10 Plant retroelement sequences and constructs which comprise the sequences are provided, as are cells, seeds, embryos and plants comprising them. Preferred are isolated nucleic acid molecules, wherein said nucleic acid molecules encode at least a portion of a plant retroelement sequence and comprises a nucleic acid sequence selected from the group consisting of:

20 20 (a) a nucleic acid sequence which has more than 95% identity to a nucleic acid sequence selected from the group consisting of: SEQ ID NO 2; SEQ ID NO 5; SEQ ID NO 7; SEQ ID NO 9; SEQ ID NO 11; SEQ ID NO 13; SEQ ID NO 15; and SEQ ID NO 17, wherein said identity can be determined using the DNAsis computer program and default parameters;

25 25 (b) a nucleic acid sequence which is selected from the group consisting of: SEQ ID NO 2; SEQ ID NO 5; SEQ ID NO 7; SEQ ID NO 9; SEQ ID NO 11; SEQ ID NO 13; SEQ ID NO 15; and SEQ ID NO 17;

30 30 (c) a nucleic acid sequence which encodes an amino acid sequence which has more than 90% identity to an amino acid sequence selected from the group consisting of SEQ ID NO 4; SEQ ID NO 6; SEQ ID NO 8; SEQ ID NO 10; SEQ ID NO 12; SEQ ID NO 14; SEQ ID NO 16; SEQ ID NO 18, wherein said identity can be determined using the DNAsis computer program and default parameters;

5 (d) a nucleic acid sequence which encodes an amino acid sequence selected from the group consisting of: SEQ ID NO 4; SEQ ID NO 6; SEQ ID NO 8; SEQ ID NO 10; SEQ ID NO 12; SEQ ID NO 14; SEQ ID NO 16; and SEQ ID NO 18;

10 (e) a nucleic acid sequence which encodes an allelic variant of an amino acid sequence selected from the group consisting of: SEQ ID NO 4; SEQ ID NO 6; SEQ ID NO 8; SEQ ID NO 10; SEQ ID NO 12; SEQ ID NO 14; SEQ ID NO 16; and SEQ ID NO 18; and

15 (f) a nucleic acid sequence fully complementary to a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); a nucleic acid sequence of (b); a nucleic acid sequence of (c); a nucleic acid sequence of (d); and a nucleic acid sequence of (e).

20 Nucleic acid molecule as above, which further comprises at least one nucleic acid sequence which encodes at least one agronomically-significant characteristic are preferred. More preferred are those nucleic acid molecules as described wherein the agronomically-significant characteristic is selected from the group consisting of: male sterility; self-incompatibility; foreign organism resistance; improved biosynthetic pathways; environmental tolerance; photosynthetic pathways; and nutrient content and those wherein the agronomically significant characteristic is selected from the group consisting of: fruit ripening; oil biosynthesis; pigment biosynthesis; seed formation; starch metabolism; salt tolerance; cold/frost tolerance; drought tolerance; tolerance to anaerobic conditions; protein content; carbohydrate content (including sugars and starches); amino acid content; and fatty acid content.

25 Seeds and plants comprising a nucleic acid molecule as described are also preferred. More preferred are plants as described, wherein the plant is selected from the group consisting of: soybean; maize; sugar cane; beet; tobacco; wheat; barley; poppy; rape; sunflower; alfalfa; sorghum; rose; carnation; gerbera; carrot; tomato; lettuce; chicory; pepper; melon; cabbage; oat; rye; cotton; flax; potato; pine; walnut; citrus (including oranges, grapefruit etc.); hemp; oak; rice; petunia; orchids; *Arabidopsis*; broccoli; cauliflower; brussel sprouts; onion; garlic; leek; squash; pumpkin; celery; pea; bean (including various legumes); strawberries; grapes; apples; pears; peaches; banana; palm; cocoa; cucumber; pineapple; apricot; plum; sugar beet; lawn grasses;

maple; triticale; safflower; peanut; and olive. Most preferred are plants as described which are soybean plants.

5 Plant retroelements comprising an amino acid sequence encoded by a nucleic acid sequence described are also provided. Plant cells comprising a nucleic acid molecule described herein, as well as plant retroviral proteins encoded by nucleic acid molecules described herein are provided.

10 Moreover, methods to transfer nucleic acid into a plant cell, comprising contacting a nucleic acid molecule of the present invention with at least one plant cell under conditions sufficient to allow said nucleic acid molecule to enter at least one cell of said plant are provided. In particular there is provided, methods to impart agronomically-significant characteristics to at least one plant cell, comprising: contacting a plant retroelement of the present invention to at least one plant cell under conditions sufficient to allow a nucleic acid molecule to enter said cell, wherein said nucleic acid molecule encodes an agronomically-significant characteristic. Methods as described, wherein the agronomically-significant characteristic is selected from the group consisting of: male sterility; self-incompatibility; foreign organism resistance; improved biosynthetic pathways; environmental tolerance; photosynthetic pathways; and nutrient content and those wherein the agronomically significant characteristic is selected from the group consisting of: fruit ripening; oil biosynthesis; pigment biosynthesis; seed formation; starch metabolism; salt tolerance; cold/frost tolerance; drought tolerance; tolerance to anaerobic conditions; protein content; carbohydrate content (including sugars and starches); amino acid content; and fatty acid content.

15

20

25

30 Plant retroelement sequences comprising specialized signals, and constructs which comprise the sequences are provided, as are cells, seeds, embryos and plants comprising them. Preferred are isolated nucleic acid molecules, comprising a nucleic acid sequence selected from the group consisting of:

35

- (a) a nucleic acid sequence which has more than 95% identity to SEQ ID NO 2; wherein said identity can be determined using the DNAsis computer program and default parameters;
- (b) a nucleic acid sequence which is SEQ ID NO 2;

5 (c) a nucleic acid sequence which encodes amino acid sequence SEQ ID NO 4; and

(d) a nucleic acid sequence fully complementary to a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); a nucleic acid sequence of (b); and a nucleic acid sequence of (c).

10 Plant retroelements as described above, which further comprise at least one nucleic acid sequence which encodes at least one agronomically-significant characteristic are preferred. More preferred are those methods wherein the agronomically-significant characteristic is selected from the group consisting of: male sterility; self-incompatibility; foreign organism resistance; improved biosynthetic pathways; environmental tolerance; photosynthetic pathways; and nutrient content and those wherein the agronomically significant characteristic is selected from the group consisting of: fruit ripening; oil biosynthesis; pigment biosynthesis; seed formation; starch metabolism; salt tolerance; cold/frost tolerance; drought tolerance; 15 tolerance to anaerobic conditions; protein content; carbohydrate content (including sugars and starches); amino acid content; and fatty acid content.

20 Preferred are plant retroviral particles comprising an isolated retroelement as described, and seeds and plants comprising the retroelements as described. More preferred plants include soybean; maize; sugar cane; beet; tobacco; wheat; barley; poppy; rape; sunflower; alfalfa; sorghum; rose; carnation; gerbera; carrot; tomato; lettuce; chicory; pepper; melon; cabbage; oat; rye; cotton; flax; potato; pine; walnut; 25 citrus (including oranges, grapefruit etc.); hemp; oak; rice; petunia; orchids; Arabidopsis; broccoli; cauliflower; brussel sprouts; onion; garlic; leek; squash; pumpkin; celery; pea; bean (including various legumes); strawberries; grapes; apples; pears; peaches; banana; palm; cocoa; cucumber; pineapple; apricot; plum; sugar beet; lawn grasses; maple; triticale; safflower; peanut; and olive. Soybean is most preferred.

30 Also provided are methods to transfer nucleic acid into a plant cell, comprising contacting a plant retroelement as described with at least one plant cell under conditions sufficient to allow said plant retroelement to enter said cell. Methods to impart agronomically-significant characteristics to a plant, comprising contacting a plant retroelement as described with at least one plant cell under conditions sufficient to allow said plant retroelement to enter said cell are also preferred.

Those methods wherein the plant retroelement is contacted with said cell via a plant retroviral particle described herein are preferred.

5 Plant retroviruses are also provided. In particular, plant retroviral particles comprising a plant-derived retrovirus envelope protein are provided. Plant retroviral particles comprising a plant-derived retrovirus envelope protein and which further comprise a plant retroviral protein selected from the group consisting of: plant-derived integrase; plant derived reverse transcriptase; plant-derived gag; and plant-derived RNaseH are preferred.

10 Plant retroviral particles comprising specialized retroviral proteins, and cells, seeds, embryos and plants which comprise the retroviral particles are provided. Preferred are isolated retroviral particles comprising a plant retroviral protein encoded by a nucleic acid sequence selected from the group consisting of:

15 (a) a nucleic acid sequence comprising (i) a nucleic acid sequence which encodes at least one plant retroviral envelope protein, and (ii) a nucleic acid sequence which has more than 60% identity to a nucleic acid sequence selected from the group consisting of: SEQ ID NO 9; SEQ ID NO 11; SEQ ID NO 15; SEQ ID NO 26; SEQ 20 ID NO 27; SEQ ID NO 28; SEQ ID NO 29; SEQ ID NO 30; and SEQ ID NO 31, wherein said identity can be determined using the DNAsis computer program and default parameters;

25 (b) a nucleic acid sequence which encodes an amino acid sequence encoded by a nucleic acid sequence (a);

(c) a nucleic acid sequence which encodes an allelic variant of an amino acid sequence encoded by a nucleic acid sequence of (a); and

30 (d) a nucleic acid sequence fully complementary to a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); a nucleic acid sequence of (b); and a nucleic acid sequence of (c).

35 In particular, there are provided plant retroviral particles, wherein said nucleic acid sequence as described in (a) comprises a plant envelope nucleic acid specifically mentioned in claim 6 is preferred. Those particles which further comprise at least

one nucleic acid sequence which encodes at least one agronomically-significant characteristic are preferred.

Also provided are methods to transfer nucleic acid into a plant cell, comprising
5 contacting a plant retroviral particle as described above to at least one plant cell under conditions sufficient to allow said nucleic acid to enter said cell. More preferred are methods to impart agronomically-significant characteristics to a plant, comprising contacting a plant retroviral particle as described to at least one plant cell under conditions sufficient to allow said nucleic acid to enter said cell.

10 More preferred are isolated retroviral particles comprising a plant retroviral protein encoded by a nucleic acid sequence selected from the group consisting of:

15 (a) a nucleic acid sequence which has more than 80% identity to a nucleic acid sequence selected from the group consisting of: SEQ ID NO 9; SEQ ID NO 11; and SEQ ID NO 15, wherein said identity can be determined using the DNAsis computer program and default parameters;

20 (b) a nucleic acid sequence which encodes a nucleic acid selected from the group consisting of: SEQ ID NO 9; SEQ ID NO 11; and SEQ ID NO 15;

25 (c) a nucleic acid sequence which encodes an amino acid sequence encoded by a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); and a nucleic acid sequence of (b);

(d) a nucleic acid sequence which encodes an allelic variant of an amino acid sequence encoded by a nucleic acid selected from the group consisting of: a nucleic acid sequence of (a); and a nucleic acid sequence of (b); and

30 (e) a nucleic acid sequence fully complementary to a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); a nucleic acid sequence of (b); a nucleic acid sequence of (c); and a nucleic acid sequence of (d).

35 Nucleic acids as above, which further comprises at least one nucleic acid sequence which encodes at least one agronomically-significant characteristic are preferred. More preferred are those nucleic acids wherein the agronomically-significant

characteristic is selected from the group consisting of: male sterility; self-incompatibility; foreign organism resistance; improved biosynthetic pathways; environmental tolerance; photosynthetic pathways; and nutrient content. Also more preferred are those isolated nucleic acid molecule as described, wherein the agronomically significant characteristic is selected from the group consisting of: fruit ripening; oil biosynthesis; pigment biosynthesis; seed formation; starch metabolism; salt tolerance; cold/frost tolerance; drought tolerance; tolerance to anaerobic conditions; protein content; carbohydrate content (including sugars and starches); amino acid content; and fatty acid content.

Also provided are methods to transfer nucleic acid into a plant cell, comprising contacting a plant retroviral particle as described above to at least one plant cell under conditions sufficient to allow said nucleic acid to enter said cell. More preferred are methods to impart agronomically-significant characteristics to a plant, comprising contacting a plant retroviral particle as described to at least one plant cell under conditions sufficient to allow said nucleic acid to enter said cell.

Also preferred are isolated retroviral particles comprising a plant retroviral protein encoded by a nucleic acid sequence selected from the group consisting of:

(a) a nucleic acid sequence which has more than 60% identity to a nucleic acid sequence selected from the group consisting of SEQ ID NO 9; SEQ ID NO 11; SEQ ID NO 15; SEQ ID NO 26; SEQ ID NO 27; SEQ ID NO 28; SEQ ID NO 29; SEQ ID NO 30; and SEQ ID NO 31, wherein said identity can be determined using the DNAsis computer program and default parameters;

(b) a nucleic acid sequence which encodes a nucleic acid selected from the group consisting of: SEQ ID NO 9; SEQ ID NO 11; SEQ ID NO 15; SEQ ID NO 26; SEQ ID NO 27; SEQ ID NO 28; SEQ ID NO 29; SEQ ID NO 30; and SEQ ID NO 31;

(c) a nucleic acid sequence which encodes an amino acid sequence encoded by a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); and a nucleic acid sequence of (b);

(d) a nucleic acid sequence which encodes an allelic variant of an amino acid sequence encoded by a nucleic acid selected from the group consisting of: a nucleic acid sequence of (a); and a nucleic acid sequence of (b); and

5 (e) a nucleic acid sequence fully complementary to a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); a nucleic acid sequence of (b); a nucleic acid sequence of (c); and a nucleic acid sequence of (d).

10 Plant retroviral particles as described above, which further comprises an envelope-encoding nucleic acid sequence specifically described herein are preferred. Preferred are those retroviral particles which further comprise at least one nucleic acid sequence which encodes at least one agronomically-significant characteristic.

15 Also provided are methods to transfer nucleic acid into a plant cell, comprising contacting a plant retroviral particle as described above to at least one plant cell under conditions sufficient to allow said nucleic acid to enter said cell. More preferred are methods to impart agronomically-significant characteristics to a plant, comprising contacting a plant retroviral particle as described to at least one plant cell under conditions sufficient to allow said nucleic acid to enter said cell.

20 “Allelic variant” is meant to refer to a full length gene or partial sequence of a full length gene that occurs at essentially the same locus (or loci) as the referent sequence, but which, due to natural variations caused by, for example, mutation or recombination, has a similar but not identical sequence. Allelic variants typically encode proteins having similar activity to that of the protein encoded by the gene to which they are being compared. Allelic variants can also comprise alterations in the 5' or 3' untranslated regions of the gene (e.g., in regulatory control regions).

25 By “agronomically-significant” it is meant any trait of a plant which is recognized by members of the agricultural industry as desirable.

30 “Fragment” is meant to refer to any subset of the referent nucleic acid molecule.

35 By “plant” it is meant one or more plant seed, plant embryo, plant part or whole plant. The plant may be an angiosperm (monocot or dicot), gymnosperm, man-made or naturally-occurring.

By "proteins" it is meant any compounds which comprise amino acids, including peptides, polypeptides, fusion proteins, etc.

5 Moreover, for the purposes of the present invention, the term "a" or "an" entity refers to one or more of that entity; for example, "a protein" or "a nucleic acid molecule" refers to one or more of those compounds or at least one compound. As such, the terms "a" (or "an"), "one or more" and "at least one" can be used interchangeably herein. It is also to be noted that the terms "comprising", "including", and "having" can be used interchangeably. Furthermore, a compound 10 "selected from the group consisting of" refers to one or more of the compounds in the list that follows, including mixtures (i.e., combinations) of two or more of the compounds. According to the present invention, an isolated, or biologically pure, 15 protein or nucleic acid molecule is a compound that has been removed from its natural milieu. As such, "isolated" and "biologically pure" do not necessarily reflect the extent to which the compound has been purified. An isolated compound of the present invention can be obtained from its natural source, can be produced 20 using molecular biology techniques or can be produced by chemical synthesis. Lastly, "more than" and "greater than" are interchangeable, and when used to modify a percent identity, ie. "more than 90% identity", mean any increment to 100%, so long as the increment were greater than the percentage specifically named. In the example of "more than 90% identity", the term would include, 25 among all other possibilities, 90.00001, 93.7, 98.04 and 99.0827 and 100%.

25 The following is a summary of the sequence listing, as a convenient reference.

SEQ ID NO	Description
1	specialized primer binding site version 1
2	specialized primer binding site version 2
3	specialized polypurine tract
4	targeting sequence
5	NA generic envelope
6	AA of 5
7	NA of generic gag
8	AA of 7
9	NA of generic integrase

10	AA of 9
11	NA of generic reverse transcriptase
12	AA of 11
13	generic protease
14	AA of 13
15	generic RNaseH
16	AA of 15
17	generic retroelement
18	AA of 17
19	NA calypso 1-1
20	NA calypso 1-2
21	NA calypso 1-3
22	NA calypso 2-1
23	NA calypso 2-2
24	NA athila env
25	NA cyclops env
26	NA athila integrase
27	NA athila reverse transcriptase
28	NA athila RNaseH
29	NA cyclops reverse transcriptase
30	NA cyclops RNaseH
31	NA cyclops integrase
32	NA calypso envelope
33	NA calypso RNaseH
34	NA calypso reverse transcriptase
35	NA calypso integrase
36	Primer binding site A
37	Primer binding site B
38	Primer binding site minimum
39	Primer binding site extended
40	polypurine tract A
41	polypurine tract B

DETAILED DESCRIPTION OF THE INVENTION

5 In one embodiment, there are provided isolated nucleic acid molecules, wherein said nucleic acid molecules encode at least a portion of a plant retroelement and comprises a nucleic acid sequence selected from the group consisting of:

10 (a) a nucleic acid sequence which is a plant retroelement primer binding site and which has more than 95% identity to SEQ ID NO 2, wherein said identity can be determined using the DNAsis computer program and default parameters;

15 (b) a nucleic acid sequence which is at least a portion of a plant retroelement envelope sequence and which has more than 50% identity to SEQ ID NO 5, wherein said identity can be determined using the DNAsis computer program and default parameters;

20 (c) a nucleic acid sequence which is at least a portion of a plant retroelement gag sequence and which has more than 50% identity to SEQ ID NO 7, wherein said identity can be determined using the DNAsis computer program and default parameters;

25 (d) a nucleic acid sequence which is at least a portion of a plant retroelement integrase sequence and which has more than 70% identity to SEQ ID NO 9, wherein said identity can be determined using the DNAsis computer program and default parameters;

30 (e) a nucleic acid sequence which is at least a portion of a plant retroelement reverse transcriptase sequence and which has more than 70% identity to SEQ ID NO 11, wherein said identity can be determined using the DNAsis computer program and default parameters;

35 (f) a nucleic acid sequence which is at least a portion of a plant retroelement protease sequence and which has more than 50% identity to SEQ ID NO 13, wherein said identity can be determined using the DNAsis computer program and default parameters;

5 (g) a nucleic acid sequence which is at least a portion of a plant retroelement RNaseH sequence and which has more than 70% identity to SEQ ID NO 15, wherein said identity can be determined using the DNAsis computer program and default parameters;

10 (h) a nucleic acid sequence which is at least a portion of a plant retroelement sequence and which has more than 50% identity to SEQ ID NO 17, wherein said identity can be determined using the DNAsis computer program and default parameters;

15 (i) a nucleic acid sequence which is selected from the group consisting of: SEQ ID NO 2; SEQ ID NO 5; SEQ ID NO 7; SEQ ID NO 9; SEQ ID NO 11; SEQ ID NO 13; SEQ ID NO 15; and SEQ ID NO 17.

20 (j) a nucleic acid sequence which encodes an amino acid sequence which is at least a portion of a plant retroelement envelope sequence and has more than 30% identity to SEQ ID NO 6, wherein said identity can be determined using the DNAsis computer program and default parameters;

25 (k) a nucleic acid sequence which encodes an amino acid sequence which is at least a portion of a plant retroelement gag sequence and has more than 30% identity to SEQ ID NO 8, wherein said identity can be determined using the DNAsis computer program and default parameters;

30 (l) a nucleic acid sequence which encodes an amino acid sequence which is at least a portion of a plant retroelement integrase sequence and has more than 75% identity to SEQ ID NO 10, wherein said identity can be determined using the DNAsis computer program and default parameters;

35 (m) a nucleic acid sequence which encodes an amino acid sequence which is at least a portion of a plant retroelement reverse transcriptase sequence and has more than 79% identity to SEQ ID NO 12, wherein said identity can be determined using the DNAsis computer program and default parameters;

5 (n) a nucleic acid sequence which encodes an amino acid sequence which is at least a portion of a plant retroelement protease sequence and has more than 55% identity to SEQ ID NO 14, wherein said identity can be determined using the DNAsis computer program and default parameters;

10 (o) a nucleic acid sequence which encodes an amino acid sequence which is at least a portion of a plant retroelement RNaseH sequence and has more than 90% identity to SEQ ID NO 16, wherein said identity can be determined using the DNAsis computer program and default parameters;

15 (p) a nucleic acid sequence which encodes an amino acid sequence which is at least a portion of a plant retroelement sequence and has more than 40% identity to SEQ ID NO 18, wherein said identity can be determined using the DNAsis computer program;

20 (q) a nucleic acid sequence which encodes an amino acid sequence selected from the group consisting of: SEQ ID NO 4; SEQ ID NO 6; SEQ ID NO 8; SEQ ID NO 10; SEQ ID NO 12; SEQ ID NO 14; SEQ ID NO 16; and SEQ ID NO 18;

25 (r) a nucleic acid sequence which encodes an allelic variant of an amino acid sequence selected from the group consisting of: SEQ ID NO 4; SEQ ID NO 6; SEQ ID NO 8; SEQ ID NO 10; SEQ ID NO 12; SEQ ID NO 14; SEQ ID NO 16; and SEQ ID NO 18; and

30 (s) a nucleic acid sequence fully complementary to a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); a nucleic acid sequence of (b); a nucleic acid sequence of (c); a nucleic acid sequence of (d); a nucleic acid sequence of (e); a nucleic acid sequence of (f); a nucleic acid sequence of (g); a nucleic acid sequence of (h); a nucleic acid sequence of (i); a nucleic acid sequence of (j); a nucleic acid sequence of (k); a nucleic acid sequence of (l); a nucleic acid sequence of (m); a nucleic acid sequence of (n); a nucleic acid sequence of (o); a nucleic acid sequence of (p); a nucleic acid sequence of (q); and a nucleic acid sequence of (r).

35 Seeds and plants comprising a nucleic acid as above are particularly provided. Nucleic acid molecules as above which comprise gag, pol and env genes and which

comprise adenine-thymidine-guanidine as the gag gene start codon are also particularly provided. Those which comprise gag, pol and env genes, the adenine-thymidine-guanidine as the gag gene start codon, and which further comprises SEQ ID NO 4 are also provided.

5

Included within the scope of the present invention, with particular regard to the nucleic acids above, are allelic variants, degenerate sequences and homologues. The present invention also includes variants due to laboratory manipulation, such as, but not limited to, variants produced during polymerase chain reaction amplification or site directed mutagenesis. It is also well known that there is a substantial amount of redundancy in the various codons which code for specific amino acids. Therefore, this invention is also directed to those nucleic acid sequences which contain alternative codons which code for the eventual translation of the identical amino acid. Also included within the scope of this invention are mutations either in the nucleic acid sequence or the translated protein which do not substantially alter the ultimate physical properties of the expressed protein. For example, substitution of valine for leucine, arginine for lysine, or asparagine for glutamine may not cause a change in functionality of the polypeptide. Lastly, a nucleic acid sequence homologous to the exemplified nucleic acid molecules (or allelic variants or degenerates thereof) will have at least 85%, preferably 90%, and most preferably 95% sequence identity with a nucleic acid molecule in the sequence listing.

25 It is known in the art that there are commercially available computer
programs for determining the degree of similarity between two nucleic acid
sequences. These computer programs include various known methods to determine
the percentage identity and the number and length of gaps between hybrid nucleic
acid molecules. Preferred methods to determine the percent identity among amino
acid sequences and also among nucleic acid sequences include analysis using one or
30 more of the commercially available computer programs designed to compare and
analyze nucleic acid or amino acid sequences. These computer programs include,
but are not limited to, GCG™ (available from Genetics Computer Group, Madison,
WI), DNAsis™ (available from Hitachi Software, San Bruno, CA) and
MacVector™ (available from the Eastman Kodak Company, New Haven, CT). A
35 preferred method to determine percent identity among amino acid sequences and
also among nucleic acid sequences includes using the Compare function by

maximum matching within the program DNAsis Version 2.1 using default parameters.

Knowing the nucleic acid sequences of the present invention allows one skilled in the art to, for example, (a) make copies of those nucleic acid molecules, (b) obtain nucleic acid molecules including at least a portion of such nucleic acid molecules (e.g., nucleic acid molecules including full-length genes, full-length coding regions, regulatory control sequences, truncated coding regions), and (c) obtain similar nucleic acid molecules from other species. Such nucleic acid molecules can be obtained in a variety of ways including screening appropriate expression libraries with antibodies of the present invention; traditional cloning techniques using oligonucleotide probes of the present invention to screen appropriate libraries of DNA; and PCR amplification of appropriate libraries of DNA using oligonucleotide primers of the present invention. Preferred libraries to screen or from which to amplify nucleic acid molecules include canine cDNA libraries as well as genomic DNA libraries. Similarly, preferred DNA sources to screen or from which to amplify nucleic acid molecules include adult cDNA and genomic DNA. Techniques to clone and amplify genes are disclosed, for example, in Sambrook et al., *ibid.*

20 Recombination constructs can be made using the starting materials above or with additional materials, using methods well-known in the art. In general, the sequences can be manipulated to have ligase-compatible ends, and incubated with ligase to generate full constructs. For example, restriction enzymes can be chosen on the basis of their ability to cut at an acceptable site in both sequence to be ligated, or a linker may be added to convert the sequence end(s) to ones that are compatible. The methods for conducting these types of molecular manipulations are well-known in the art, and are described in detail in Sambrook et al., Molecular Cloning. A Laboratory Manual (Cold Spring Harbor Laboratory Press, 1989) and Ausubel et al., Current Protocols in Molecular Biology (Greene Publishing Associates, Inc., 1993). The methods described herein according to Tinland et al., 91 Proc. Natl. Acad. Sci. USA 8000 (1994) can also be used.

35 The present invention also includes nucleic acid molecules that are oligonucleotides capable of hybridizing, under stringent hybridization conditions, with complementary regions of other, preferably longer, nucleic acid molecules of

the present invention. Oligonucleotides of the present invention can be RNA, DNA, or derivatives of either. The minimum size of such oligonucleotides is the size required for formation of a stable hybrid between an oligonucleotide and a complementary sequence on a nucleic acid molecule of the present invention. 5 Minimal size characteristics are disclosed herein. The present invention includes oligonucleotides that can be used as, for example, probes to identify nucleic acid molecules, primers to produce nucleic acid molecules or therapeutic reagents. Stringent hybridization conditions are determined based on defined physical properties of the gene to which the nucleic acid molecule is being hybridized, and can be defined mathematically. Stringent hybridization conditions are those 10 experimental parameters that allow an individual skilled in the art to identify significant similarities between heterologous nucleic acid molecules. These conditions are well known to those skilled in the art. See, for example, Sambrook, et al., 1989, Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Labs 15 Press, and Meinkoth, et al., 1984, Anal. Biochem. 138, 267-284.

Recombinant molecules of the present invention may also (a) contain 20 secretory signals (i.e., signal segment nucleic acid sequences) to enable an expressed protein of the present invention to be secreted from the cell that produces the protein and/or (b) contain fusion sequences which lead to the expression of nucleic acid molecules of the present invention as fusion proteins. Recombinant molecules may also include intervening and/or untranslated sequences surrounding and/or within the nucleic acid sequences of nucleic acid molecules of the present invention. 25

One embodiment of the present invention includes recombinant vectors, 30 which include at least one isolated nucleic acid molecule of the present invention, inserted into any vector capable of delivering the nucleic acid molecule into a host cell. Such a vector contains heterologous nucleic acid sequences, that is nucleic acid sequences that are not naturally found adjacent to nucleic acid molecules of the present invention and that preferably are derived from a species other than the species from which the nucleic acid molecule(s) are derived. The vector can be either RNA or DNA, either prokaryotic or eukaryotic, and typically is a virus or a plasmid. Recombinant vectors can be used in the cloning, sequencing, and/or 35 otherwise manipulation of nucleic acid molecules of the present invention.

5 One type of recombinant vector, referred to herein as a recombinant molecule, comprises a nucleic acid molecule of the present invention operatively linked to an expression vector. The phrase operatively linked refers to insertion of a nucleic acid molecule into an expression vector in a manner such that the molecule is able to be expressed when transformed into a host cell. As used herein, an expression vector is a DNA or RNA vector that is capable of transforming a host cell and of effecting expression of a specified nucleic acid molecule. Expression vectors can be either prokaryotic or eukaryotic, and are typically viruses or plasmids. Expression vectors of the present invention include any vectors that function (i.e., direct gene expression) in recombinant cells of the present invention, including in bacterial, fungal, endoparasite, insect, other animal, and plant cells.

10

15 In particular, expression vectors of the present invention contain regulatory sequences such as transcription control sequences, translation control sequences, origins of replication, and other regulatory sequences that are compatible with the recombinant cell and that control the expression of nucleic acid molecules of the present invention. In particular, recombinant molecules of the present invention include transcription control sequences. Transcription control sequences are sequences which control the initiation, elongation, and termination of transcription.

20 Particularly important transcription control sequences are those which control transcription initiation, such as promoter, enhancer, operator and repressor sequences. Suitable transcription control sequences include any transcription control sequences that can function in at least one of the recombinant cells of the present invention. A variety of such transcription control sequences are known to those skilled in the art. Preferred transcription control sequences include those which function in bacterial, yeast, insect and mammalian cells, such as, but not limited to, tac, lac, trp, trc, oxy-pro, omp/lpp, rrnB, bacteriophage lambda (such as lambda pL and lambda pR and fusions that include such promoters), bacteriophage T7, T7lac, bacteriophage T3, bacteriophage SP6, bacteriophage SP01, metallothionein, alpha-mating factor, *Pichia* alcohol oxidase, alphavirus subgenomic promoters (such as *Sindbis* virus subgenomic promoters), antibiotic resistance gene, baculovirus, *Heliothis zea* insect virus, *vaccinia* virus, herpesvirus, raccoon poxvirus, other poxvirus, adenovirus, cytomegalovirus (such as intermediate early promoters), simian virus 40, retrovirus, actin, retroviral long terminal repeat, Rous sarcoma virus, heat shock, phosphate and nitrate transcription control sequences as well as other sequences capable of controlling gene expression

25

30

35

5 in prokaryotic or eukaryotic cells. Additional suitable transcription control sequences include tissue-specific promoters and enhancers as well as lymphokine-inducible promoters (e.g., promoters inducible by interferons or interleukins). Transcription control sequences of the present invention can also include naturally occurring transcription control sequences naturally associated with plants. The present invention also comprises expression vectors comprising a nucleic acid molecule described herein.

10 For instance, the following promoters would be useful in early expression of the present sequences: Ogs4B (Tsuchiya et al., 36 Plant Cell Physiology 487 (1994); TA29 (Koltunow et al., 2 Plant Cell 1201 (1990); A3 & A9 (Paul et al., 19 Plant Molecular Biology 611 (1992). In order to then constitutively express the sequences described above, the construct optionally contains, for example, a 35S promoter.

15 Vectors which comprise the above sequences are within the scope of the present invention, as are plants transformed with the above sequences. Vectors may be obtained from various commercial sources, including Clontech Laboratories, Inc. (Palo Alto, CA), Stratagene (La Jolla, CA), Invitrogen (Carlsbad, CA), New England Biolabs (Beverly, MA) and Promega (Madison, WI). Preferred vectors are those which are capable of transferring the sequences disclosed herein into plant cells or plant parts.

20 Recombinant DNA technologies can be used to improve expression of transformed nucleic acid molecules by manipulating, for example, the number of copies of the nucleic acid molecules within a host cell, the efficiency with which those nucleic acid molecules are transcribed, the efficiency with which the resultant transcripts are translated, and the efficiency of post-translational modifications. Preferred vectors are those which are capable of transferring the sequences disclosed herein into plant cells or plant parts.

25 Recombinant techniques useful for increasing the expression of nucleic acid molecules of the present invention include, but are not limited to, operatively linking nucleic acid molecules to high-copy number plasmids, integration of the nucleic acid molecules into one or more host cell chromosomes, addition of vector stability sequences to plasmids, substitutions or modifications of transcription control signals (e.g., promoters, operators, enhancers), substitutions or modifications of translational control signals (e.g., ribosome binding sites, Shine-Dalgarno sequences), modification of nucleic acid molecules of the present

30

35

invention to correspond to the codon usage of the host cell, deletion of sequences that destabilize transcripts, and use of control signals that temporally separate recombinant cell growth from recombinant enzyme production during fermentation. The activity of an expressed recombinant protein of the present invention may be improved by fragmenting, modifying, or derivatizing nucleic acid molecules encoding such a protein.

Nucleic acids of the present invention may be transferred to cells according to the methods of the present invention, as well as using any of the following well-known means: infective, vector-containing bacterial strains (such as *Agrobacterium rhizogenes* and *Agrobacterium tumefaciens*) according to ie. Zambryski, 43 Ann. Rev. Pl. Physiol. Pl. Mol. Biol. 465 (1992); pollen-tube transformation [Zhong-xun et al., 6 Plant Molec. Bio. 165 (1988)]; direct transformation of germinating seeds [Toepfer et al., 1 Plant Cell 133 (1989)]; polyethylene glycol or electroporation transformation [Christou et al., 84 Proc. Nat. Acad. Sci. 3662 (1987)]; and biolistic processes [Yang & Cristou, Particle Bombardment Technology for Gene Transfer (1994)].

The transformed cells may be induced to form transformed plants via organogenesis or embryogenesis, according to the procedures of Dixon Plant Cell Culture: A Practical Approach (IRL Press, Oxford 1987).

Any seed, embryo, plant or plant part is amenable to the present techniques. Of course, the agronomically-significant seeds, embryos, plants or plant parts are preferred. Soybean; maize; sugar cane; beet; tobacco; wheat; barley; poppy; rape; sunflower; alfalfa; sorghum; rose; carnation; gerbera; carrot; tomato; lettuce; chicory; pepper; melon; cabbage; oat; rye; cotton; flax; potato; pine; walnut; citrus (including oranges, grapefruit etc.); hemp; oak; rice; petunia; orchids; *Arabidopsis*; broccoli; cauliflower; brussel sprouts; onion; garlic; leek; squash; pumpkin; celery; pea; bean (including various legumes); strawberries; grapes; apples; pears; peaches; banana; palm; cocoa; cucumber; pineapple; apricot; plum; sugar beet; lawn grasses; maple; triticale; safflower; peanut; and olive are among the preferred seeds, embryos, plants or plant parts. Particularly preferred are: soybean, tobacco and maize seeds, embryos, plants or plant parts. However, *Arabidopsis* seeds, embryos, plants or plant parts are also preferred, since it is an excellent system for study of plant genetics.

Preferred are those genes or sequences which are agronomically significant. For example, genes encoding male sterility, foreign organism resistance (viruses or bacteria), including genes which produce bacterial endotoxins, such as bacillus thuringiensis endotoxin, genes involved in specific biosynthetic pathways (eg. in fruit ripening, oil or pigment biosynthesis, seed formation, or carbohydrate metabolism), genes involved in environmental tolerance (eg. salt tolerance, lodging tolerance, cold/frost tolerance, drought tolerance, or tolerance to anaerobic conditions), or genes involved in nutrient content (eg. protein content, carbohydrate content, amino acid content, fatty acid content), genes involved in photosynthetic pathways, or genes involved in self-incompatibility. The choice of gene or sequence induced to recombine in the present invention is not limited. Examples of genes and how to obtain them are available through reference articles, books and supply catalogs, such as The Sourcebook (1-800-551-5291). Sambrook et al., Molecular Cloning. A Laboratory Manual (Cold Spring Harbor Laboratory Press, 1989) and Weising et al., 22 Ann Rev. Gen. 421 (1988) contain a synthesis of the information that is well-known in this art.

Plant envelope sequences and constructs which comprise the sequences are provided, as are cells, seeds, embryos and plants comprising them. Preferred are isolated nucleic acid molecules, wherein said nucleic acid molecules encode at least a portion of a plant envelope sequence and comprises a nucleic acid sequence selected from the group consisting of:

- (a) a nucleic acid sequence which has more than 90% identity to SEQ ID NO 5, wherein said identity can be determined using the DNAsis computer program and default parameters;
- (b) a nucleic acid sequence which encodes SEQ ID NO 5;
- (c) a nucleic acid sequence which encodes an amino acid sequence which has greater than 85% identity to SEQ ID NO 6, wherein said identity can be determined using the DNAsis computer program and default parameters;
- (d) a nucleic acid sequence which encodes amino acid sequence SEQ ID NO 6;

5 (e) a nucleic acid sequence which encodes an allelic variant of SEQ ID NO 6; and
5 (f) a nucleic acid sequence fully complementary to a nucleic acid sequence selected
from the group consisting of: a nucleic acid sequence of (a); a nucleic acid
sequence of (b); a nucleic acid sequence of (c); a nucleic acid sequence of (d); and a
nucleic acid sequence of (e).

10 Plant cells comprising an isolated nucleic acid molecule above are particularly
preferred. Also preferred are plant envelope proteins comprising an amino acid
sequence encoded by the above. Methods to impart agronomically-significant
characteristics to at least one plant cell are also provided, comprising: contacting a
plant envelope protein as described to at least one plant cell under conditions
sufficient to allow a nucleic acid molecule to enter said cell, wherein said nucleic
acid molecule encodes an agronomically-significant characteristic.

15 Plant integrase sequences and constructs which comprise the sequences are
provided, as are cells, seeds, embryos and plants comprising them. Preferred are
isolated nucleic acid molecules, wherein said nucleic acid molecules encode at least
a portion of a plant integrase sequence and comprises a nucleic acid sequence
20 selected from the group consisting of:

25 (a) a nucleic acid sequence which has more than 90% identity to SEQ ID NO 9,
wherein said identity can be determined using the DNAsis computer program and
default parameters;

(b) a nucleic acid sequence which encodes SEQ ID NO 9;

30 (c) a nucleic acid sequence which encodes an amino acid sequence which has
greater than 85% identity to SEQ ID NO 10, wherein said identity can be
determined using the DNAsis computer program and default parameters;

(d) a nucleic acid sequence which encodes amino acid sequence SEQ ID NO 10;

(e) a nucleic acid sequence which encodes an allelic variant of SEQ ID NO 10; and

5 (f) a nucleic acid sequence fully complementary to a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); a nucleic acid sequence of (b); a nucleic acid sequence of (c); a nucleic acid sequence of (d); and a nucleic acid sequence of (e).

10 Plant cells comprising an isolated nucleic acid molecule above are particularly preferred. Also preferred are plant integrase proteins comprising an amino acid sequence encoded by the above. Methods to impart agronomically-significant characteristics to at least one plant cell are also provided, comprising: contacting a plant integrase protein as described to at least one plant cell under conditions sufficient to allow a nucleic acid molecule to enter said cell, wherein said nucleic acid molecule encodes an agronomically-significant characteristic.

15 Plant reverse transcriptase sequences and constructs which comprise the sequences are provided, as are cells, seeds, embryos and plants comprising them. Preferred are isolated nucleic acid molecules, wherein said nucleic acid molecules encode at least a portion of a plant reverse transcriptase sequence and comprises a nucleic acid sequence selected from the group consisting of:

20 (a) a nucleic acid sequence which has more than 90% identity to SEQ ID NO 11, wherein said identity can be determined using the DNAsis computer program and default parameters;

25 (b) a nucleic acid sequence which encodes SEQ ID NO 11;

(c) a nucleic acid sequence which encodes an amino acid sequence which has greater than 85% identity to SEQ ID NO 12, wherein said identity can be determined using the DNAsis computer program and default parameters;

30 (d) a nucleic acid sequence which encodes amino acid sequence SEQ ID NO 12;

(e) a nucleic acid sequence which encodes an allelic variant of SEQ ID NO 12; and

5 (f) a nucleic acid sequence fully complementary to a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); a nucleic acid sequence of (b); a nucleic acid sequence of (c); a nucleic acid sequence of (d); and a nucleic acid sequence of (e).

10 Plant cells comprising an isolated nucleic acid molecule above are particularly preferred. Also preferred are plant reverse transcriptase proteins comprising an amino acid sequence encoded by the above. Methods to impart agronomically-significant characteristics to at least one plant cell are also provided, comprising: contacting a plant reverse transcriptase protein as described to at least one plant cell under conditions sufficient to allow a nucleic acid molecule to enter said cell, wherein said nucleic acid molecule encodes an agronomically-significant characteristic.

15 Plant RNaseH sequences and constructs which comprise the sequences are provided, as are cells, seeds, embryos and plants comprising them. Preferred are isolated nucleic acid molecules, wherein said nucleic acid molecules encode at least a portion of a plant RNaseH sequence and comprises a nucleic acid sequence selected from the group consisting of:

20 (a) a nucleic acid sequence which has more than 90% identity to SEQ ID NO 15, wherein said identity can be determined using the DNAsis computer program and default parameters;

25 (b) a nucleic acid sequence which encodes SEQ ID NO 15;

(c) a nucleic acid sequence which encodes an amino acid sequence which has greater than 95% identity to SEQ ID NO 16, wherein said identity can be determined using the DNAsis computer program and default parameters;

30 (d) a nucleic acid sequence which encodes amino acid sequence SEQ ID NO 16;

(e) a nucleic acid sequence which encodes an allelic variant of SEQ ID NO 16; and

35 (f) a nucleic acid sequence fully complementary to a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); a nucleic acid

sequence of (b); a nucleic acid sequence of (c); a nucleic acid sequence of (d); and a nucleic acid sequence of (e).

5 Plant cells comprising an isolated nucleic acid molecule above are particularly preferred. Also preferred are plant RNaseH proteins comprising an amino acid sequence encoded by the above. Methods to impart agronomically-significant characteristics to at least one plant cell are also provided, comprising: contacting a plant RNaseH protein as described to at least one plant cell under conditions sufficient to allow a nucleic acid molecule to enter said cell, wherein said nucleic acid molecule encodes an agronomically-significant characteristic.

10 15 Plant retroelement sequences and constructs which comprise the sequences are provided, as are cells, seeds, embryos and plants comprising them. Preferred are isolated nucleic acid molecules, wherein said nucleic acid molecules encode at least a portion of a plant retroelement sequence and comprises a nucleic acid sequence selected from the group consisting of:

20 (a) a nucleic acid sequence which has more than 95% identity to a nucleic acid sequence selected from the group consisting of: SEQ ID NO 2; SEQ ID NO 5; SEQ ID NO 7; SEQ ID NO 9; SEQ ID NO 11; SEQ ID NO 13; SEQ ID NO 15; and SEQ ID NO 17, wherein said identity can be determined using the DNAsis computer program and default parameters;

25 (b) a nucleic acid sequence which is selected from the group consisting of: SEQ ID NO 2; SEQ ID NO 5; SEQ ID NO 7; SEQ ID NO 9; SEQ ID NO 11; SEQ ID NO 13; SEQ ID NO 15; and SEQ ID NO 17;

30 (c) a nucleic acid sequence which encodes an amino acid sequence which has more than 90% identity to an amino acid sequence selected from the group consisting of SEQ ID NO 4; SEQ ID NO 6; SEQ ID NO 8; SEQ ID NO 10; SEQ ID NO 12; SEQ ID NO 14; SEQ ID NO 16; SEQ ID NO 18, wherein said identity can be determined using the DNAsis computer program and default parameters;

35 (d) a nucleic acid sequence which encodes an amino acid sequence selected from the group consisting of: SEQ ID NO 4; SEQ ID NO 6; SEQ ID NO 8; SEQ ID NO 10; SEQ ID NO 12; SEQ ID NO 14; SEQ ID NO 16; and SEQ ID NO 18;

5 (e) a nucleic acid sequence which encodes an allelic variant of an amino acid sequence selected from the group consisting of: SEQ ID NO 4; SEQ ID NO 6; SEQ ID NO 8; SEQ ID NO 10; SEQ ID NO 12; SEQ ID NO 14; SEQ ID NO 16; and SEQ ID NO 18; and

10 (f) a nucleic acid sequence fully complementary to a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); a nucleic acid sequence of (b); a nucleic acid sequence of (c); a nucleic acid sequence of (d); and a nucleic acid sequence of (e).

15 Nucleic acid molecule as above, which further comprises at least one nucleic acid sequence which encodes at least one agronomically-significant characteristic are preferred. More preferred are those nucleic acid molecules as described wherein the agronomically-significant characteristic is selected from the group consisting of: male sterility; self-incompatibility; foreign organism resistance; improved biosynthetic pathways; environmental tolerance; photosynthetic pathways; and nutrient content. Also more preferred are those isolated nucleic acid molecule as described, wherein the agronomically significant characteristic is selected from the group consisting of: fruit ripening; oil biosynthesis; pigment biosynthesis; seed formation; starch metabolism; salt tolerance; cold/frost tolerance; drought tolerance; tolerance to anaerobic conditions; protein content; carbohydrate content (including sugars and starches); amino acid content; and fatty acid content.

20

25 Seeds and plants comprising a nucleic acid molecule as described are also preferred. More preferred are plants as described, wherein the plant is selected from the group consisting of: soybean; maize; sugar cane; beet; tobacco; wheat; barley; poppy; rape; sunflower; alfalfa; sorghum; rose; carnation; gerbera; carrot; tomato; lettuce; chicory; pepper; melon; cabbage; oat; rye; cotton; flax; potato; pine; walnut; citrus (including oranges, grapefruit etc.); hemp; oak; rice; petunia; orchids; *Arabidopsis*; broccoli; cauliflower; brussel sprouts; onion; garlic; leek; squash; pumpkin; celery; pea; bean (including various legumes); strawberries; grapes; apples; pears; peaches; banana; palm; cocoa; cucumber; pineapple; apricot; plum; sugar beet; lawn grasses; maple; triticale; safflower; peanut; and olive. Most preferred are plants as described which is a soybean plant.

30

35

Plant retroelements comprising an amino acid sequence encoded by a nucleic acid sequence described are also provided. Plant cells comprising a nucleic acid molecule described herein, as well as plant retroviral proteins encoded by nucleic acid molecules described herein are provided.

5

Moreover, methods to transfer nucleic acid into a plant cell, comprising contacting a nucleic acid molecule of the present invention with at least one plant cell under conditions sufficient to allow said nucleic acid molecule to enter at least one cell of said plant are provided. In particular there is provided, methods to impart agronomically-significant characteristics to at least one plant cell, comprising: contacting a plant retroelement of the present invention to at least one plant cell under conditions sufficient to allow a nucleic acid molecule to enter said cell, wherein said nucleic acid molecule encodes an agronomically-significant characteristic. Methods as described, wherein the agronomically-significant characteristic is selected from the group consisting of: male sterility; self-incompatibility; foreign organism resistance; improved biosynthetic pathways; environmental tolerance; photosynthetic pathways; and nutrient content are preferred, as are methods wherein the agronomically-significant characteristic is selected from the group consisting of: fruit ripening; oil biosynthesis; pigment biosynthesis; seed formation; starch metabolism; salt tolerance; cold/frost tolerance; drought tolerance; tolerance to anaerobic conditions; protein content; carbohydrate content (including sugars and starches); amino acid content; and fatty acid content.

10

15

20

25

Plant retroelement sequences comprising specialized signals, and constructs which comprise the sequences are provided, as are cells, seeds, embryos and plants comprising them. Preferred are isolated nucleic acid molecules, comprising a nucleic acid sequence selected from the group consisting of:

30

35

- (a) a nucleic acid sequence which has more than 95% identity to SEQ ID NO 2; wherein said identity can be determined using the DNAsis computer program and default parameters;
- (b) a nucleic acid sequence which is SEQ ID NO 2;
- (c) a nucleic acid sequence which encodes amino acid sequence SEQ ID NO 4; and

(d) a nucleic acid sequence fully complementary to a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); a nucleic acid sequence of (b); and a nucleic acid sequence of (c).

5 Plant retroelements as described above, which further comprise at least one nucleic acid sequence which encodes at least one agronomically-significant characteristic are preferred. More preferred are those methods wherein the agronomically-significant characteristic is selected from the group consisting of: male sterility; self-incompatibility; foreign organism resistance; improved biosynthetic pathways; 10 environmental tolerance; photosynthetic pathways; and nutrient content or those wherein the agronomically significant characteristic is selected from the group consisting of: fruit ripening; oil biosynthesis; pigment biosynthesis; seed formation; starch metabolism; salt tolerance; cold/frost tolerance; drought tolerance; tolerance to anaerobic conditions; protein content; carbohydrate content (including 15 sugars and starches); amino acid content; and fatty acid content.

Preferred are plant retroviral particles comprising an isolated retroelement as described, and seeds and plants comprising the retroelements as described. More preferred plants include soybean; maize; sugar cane; beet; tobacco; wheat; barley; 20 poppy; rape; sunflower; alfalfa; sorghum; rose; carnation; gerbera; carrot; tomato; lettuce; chicory; pepper; melon; cabbage; oat; rye; cotton; flax; potato; pine; walnut; citrus (including oranges, grapefruit etc.); hemp; oak; rice; petunia; orchids; Arabidopsis; broccoli; cauliflower; brussel sprouts; onion; garlic; leek; squash; 25 pumpkin; celery; pea; bean (including various legumes); strawberries; grapes; apples; pears; peaches; banana; palm; cocoa; cucumber; pineapple; apricot; plum; sugar beet; lawn grasses; maple; triticale; safflower; peanut; and olive. Soybean is most preferred.

30 Also provided are methods to transfer nucleic acid into a plant cell, comprising contacting a plant retroelement as described with at least one plant cell under conditions sufficient to allow said plant retroelement to enter said cell. Methods to impart agronomically-significant characteristics to a plant, comprising contacting a plant retroelement as described with at least one plant cell under conditions sufficient to allow said plant retroelement to enter said cell are also preferred. 35 Those methods wherein the plant retroelement is contacted with said cell via a plant retroviral particle described herein are preferred.

5 Plant retroviruses are also provided. In particular, plant retroviral particles comprising a plant-derived retrovirus envelope protein are provided. Plant retroviral particles comprising a plant-derived retrovirus envelope protein and which further comprise a plant retroviral protein selected from the group consisting of: plant-derived integrase; plant derived reverse transcriptase; plant-derived gag; and plant-derived RNaseH are preferred.

10 Plant retroviral particles comprising specialized retroviral proteins, and cells, seeds, embryos and plants which comprise the retroviral particles are provided. Preferred are isolated retroviral particles comprising a plant retroviral protein encoded by a nucleic acid sequence selected from the group consisting of:

15 (a) a nucleic acid sequence comprising (i) a nucleic acid sequence which encodes at least one plant retroviral envelope protein, and (ii) a nucleic acid sequence which has more than 60% identity to a nucleic acid sequence selected from the group consisting of: SEQ ID NO 9; SEQ ID NO 11; SEQ ID NO 15; SEQ ID NO 26; SEQ ID NO 27; SEQ ID NO 28; SEQ ID NO 29; SEQ ID NO 30; and SEQ ID NO 31, wherein said identity can be determined using the DNAsis computer program and default parameters;

20 (b) a nucleic acid sequence which encodes an amino acid sequence encoded by a nucleic acid sequence (a);

25 (c) a nucleic acid sequence which encodes an allelic variant of an amino acid sequence encoded by a nucleic acid sequence of (a); and

30 (d) a nucleic acid sequence fully complementary to a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); a nucleic acid sequence of (b); and a nucleic acid sequence of (c).

35 In particular, there are provided plant retroviral particles, wherein said nucleic acid sequence as described in (a) comprises a plant envelope nucleic acid specifically mentioned in claim 6 is preferred. Those particles which further comprise at least one nucleic acid sequence which encodes at least one agronomically-significant characteristic are preferred.

5 Also provided are methods to transfer nucleic acid into a plant cell, comprising contacting a plant retroviral particle as described above to at least one plant cell under conditions sufficient to allow said nucleic acid to enter said cell. More preferred are methods to impart agronomically-significant characteristics to a plant, comprising contacting a plant retroviral particle as described to at least one plant cell under conditions sufficient to allow said nucleic acid to enter said cell.

10 More preferred are isolated retroviral particles comprising a plant retroviral protein encoded by a nucleic acid sequence selected from the group consisting of:

15 (a) a nucleic acid sequence which has more than 80% identity to a nucleic acid sequence selected from the group consisting of: SEQ ID NO 9; SEQ ID NO 11; and SEQ ID NO 15, wherein said identity can be determined using the DNAsis computer program and default parameters;

20 (b) a nucleic acid sequence which encodes a nucleic acid selected from the group consisting of: SEQ ID NO 9; SEQ ID NO 11; and SEQ ID NO 15;

25 (c) a nucleic acid sequence which encodes an amino acid sequence encoded by a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); and a nucleic acid sequence of (b);

30 (d) a nucleic acid sequence which encodes an allelic variant of an amino acid sequence encoded by a nucleic acid selected from the group consisting of: a nucleic acid sequence of (a); and a nucleic acid sequence of (b); and

35 (e) a nucleic acid sequence fully complementary to a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); a nucleic acid sequence of (b); a nucleic acid sequence of (c); and a nucleic acid sequence of (d).

Nucleic acids as above, which further comprises at least one nucleic acid sequence which encodes at least one agronomically-significant characteristic are preferred. More preferred are those nucleic acids wherein the agronomically-significant characteristic is selected from the group consisting of: male sterility; self-incompatibility; foreign organism resistance; improved biosynthetic pathways;

5 environmental tolerance; photosynthetic pathways; and nutrient content, or wherein the agronomically significant characteristic is selected from the group consisting of: fruit ripening; oil biosynthesis; pigment biosynthesis; seed formation; starch metabolism; salt tolerance; cold/frost tolerance; drought tolerance; tolerance to anaerobic conditions; protein content; carbohydrate content (including sugars and starches); amino acid content; and fatty acid content.

10 Also provided are methods to transfer nucleic acid into a plant cell, comprising contacting a plant retroviral particle as described above to at least one plant cell under conditions sufficient to allow said nucleic acid to enter said cell. More preferred are methods to impart agronomically-significant characteristics to a plant, comprising contacting a plant retroviral particle as described to at least one plant cell under conditions sufficient to allow said nucleic acid to enter said cell.

15 Also preferred are isolated retroviral particles comprising a plant retroviral protein encoded by a nucleic acid sequence selected from the group consisting of:

20 (a) a nucleic acid sequence which has more than 60% identity to a nucleic acid sequence selected from the group consisting of SEQ ID NO 9; SEQ ID NO 11; SEQ ID NO 15; SEQ ID NO 26; SEQ ID NO 27; SEQ ID NO 28; SEQ ID NO 29; SEQ ID NO 30; and SEQ ID NO 31, wherein said identity can be determined using the DNAsis computer program and default parameters;

25 (b) a nucleic acid sequence which encodes a nucleic acid selected from the group consisting of: SEQ ID NO 9; SEQ ID NO 11; SEQ ID NO 15; SEQ ID NO 26; SEQ ID NO 27; SEQ ID NO 28; SEQ ID NO 29; SEQ ID NO 30; and SEQ ID NO 31;

30 (c) a nucleic acid sequence which encodes an amino acid sequence encoded by a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); and a nucleic acid sequence of (b);

(d) a nucleic acid sequence which encodes an allelic variant of an amino acid sequence encoded by a nucleic acid selected from the group consisting of: a nucleic acid sequence of (a); and a nucleic acid sequence of (b); and

(e) a nucleic acid sequence fully complementary to a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); a nucleic acid sequence of (b); a nucleic acid sequence of (c); and a nucleic acid sequence of (d).

5 Also preferred are isolated retroviral particles comprising a plant retroviral sequence encoded by a nucleic acid sequence selected from the group consisting of:

10 (a) a nucleic acid sequence which has more than 80% identity to a nucleic acid sequence selected from the group consisting of SEQ ID NO 1; SEQ ID NO 2; SEQ ID NO 3, wherein said identity can be determined using the DNAsis computer program and default parameters;

15 (b) a nucleic acid sequence which encodes a nucleic acid selected from the group consisting of: SEQ ID NO 1; SEQ ID NO 2; and SEQ ID NO 3;

(c) a nucleic acid sequence which encodes SEQ ID NO 4;

20 (d) a nucleic acid sequence which encodes an amino acid sequence encoded by a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); a nucleic acid sequence of (b); and a nucleic acid sequence of (c);

25 (e) a nucleic acid sequence which encodes an allelic variant of an amino acid sequence encoded by a nucleic acid selected from the group consisting of: a nucleic acid sequence of (a); a nucleic acid sequence of (b); and a nucleic acid sequence of (c) and

30 (f) a nucleic acid sequence fully complementary to a nucleic acid sequence selected from the group consisting of: a nucleic acid sequence of (a); a nucleic acid sequence of (b); a nucleic acid sequence of (c); a nucleic acid sequence of (e); and a nucleic acid sequence of (f).

35 Plant retroviral particles as described above, which further comprises an envelope-encoding nucleic acid sequence specifically described herein are preferred. Preferred are those retroviral particles which further comprise at least one nucleic acid sequence which encodes at least one agronomically-significant characteristic.

5 Also provided are methods to transfer nucleic acid into a plant cell, comprising contacting a plant retroviral particle as described above to at least one plant cell under conditions sufficient to allow said nucleic acid to enter said cell. More preferred are methods to impart agronomically-significant characteristics to a plant, comprising contacting a plant retroviral particle as described to at least one plant cell under conditions sufficient to allow said nucleic acid to enter said cell.

10 Also provided, as part of the present invention, are isolated nucleic acid having at least 20 contiguous nucleotides of the sequence shown in SEQ ID NO 17. "At least" means that this is the lower limit and the number can be any whole number increment up to the total number of bases in SEQ ID NO 17. For example, isolated nucleic acid sequences which are 25, 30, 35, 40, 45, 50, 55, 60, 65 and 70 are within the scope of the present invention.

15 The following paragraph is designed to elaborate on the best mode and is not indicative of the sole means for making and carrying out the present invention. This paragraph is not intended to be limiting. The best way to make the present nucleic acids is to clone the nucleic acids from the respective organisms or amplified from genomic cDNA by the polymerase chain reaction using appropriate primers. 20 The best way to make the present retroelements is to assemble the nucleic acids using standard cloning procedures. Transcriptional controls can be manipulated by inserting enhancers in or near the 5' LTR. Marker genes or genes of interest can be inserted within the retroelement. The best way to make the present retroviral particles is to express the retroelement, preferably at high levels, in plant cells and the particles harvested by sucrose gradient fractionation. The best way to use the 25 present nucleic acids is by allowing retroviral particles to come into contact with plant cells. Expression of marker genes carried by the retroelement can be used as one measure of infection and integration.

30 The following examples are not intended to limit the scope of the present invention as described and claimed. They are simply for the purpose of illustration.

EXAMPLES

Example 1 Characterizing the Arabidopsis Retroelements ("Tat" and "Athila" elements)

5

Plant material and Southern hybridizations: The Arabidopsis Information Service supplied the following seed stocks (Kranz and Kirchheim (1987) 10 Arabidopsis Inform. Serv. 24): Col-0, La-0, Kas-1, Co-4, Sei-0, Mv-0, Li-0, Cvi-0, Fi-3, Ba-1, Hau-0, Aa-0, Ms-0, Ag-0, Ge-0, No-0 and Mh-0. Genomic DNA was extracted using Qiagen genomic tips and protocols supplied by Qiagen. For 15 Southern hybridizations, the resulting DNA was digested with EcoRI, electrophoresed on 0.8% agarose and transferred to Gene Screen Plus membranes using the manufacturer's alkaline transfer protocol (New England Nuclear). All hybridizations were performed as described. Church and Gilbert (1984) Proc. Natl. Acad. Sci. USA 81: 1991-1995.

10

15

20

25

30

35

Library screening, probe preparation and PCR: Tat1 clones were obtained by screening a Landsberg erecta (La-0) 1 phage library (Voytas et al. (1990) Genetics 126: 713-721), using a probe derived by PCR amplification of La-0 DNA. The primers for probe amplification were based on the three published Tat1 20 sequences (DVO158, 5'-GGGATCCGCAATTAGAATCT-3'; DVO159, 5'-CGAATTCTGGTCCACTTCGGA-3'). Peleman et al. (1991) Proc. Natl. Acad. Sci. USA 88: 3618-3622. Subsequent probes were restriction fragments of cloned Tat1 elements, and all probes were radiolabeled by random priming (Promega). Long PCR was performed using the Expand Long Template PCR System (Boehringer Mannheim) with LTR-specific primers (DVO354, 5'-CCACAAGATTCTAATTGCGGATTC-3'; DVO355, 5'-CCGAAATGGACCGAACCCGACATC-3'). The protocol used was for PCR 25 amplification of DNA up to 15 kb. The following PCR primers were used to confirm the structure of Tat1-3: DVO405 (5'-TTTCCAGGCTTTGACGAGATTG-3') for the 3' non-coding region, DVO385 (5'-CGACTCGAGCTCCATAGCGATG-3') for the second ORF of Tat1-3 (note that the seventh base was changed from an A to a G to make an XhoI and a SalI restriction site) and DVO371 (5'-CGGATTGGGCCGAAATGGACCGAA-3') for the 3' LTR.

5 DNA sequencing: Clones were sequenced either by the DNA sequencing facility at Iowa State University or with the fmol sequencing kit (Promega). DNA from the λ phage clones was initially subcloned into the vector pBluescript II KS- and transformed into the *E. coli* host strain XL1 Blue (Stratagene). AUSUBEL et al. (1987) Current Protocols in Molecular Biology. Greene/Wiley Interscience, New York. Subclones in the vector pMOB were used for transposon mutagenesis with the TN 1000 sequencing kit (Gold Biotechnologies). Transposon-specific primers were used for DNA sequencing reactions.

10 Sequence analysis: Sequence analysis was performed using the GCG software package (Devereux et al. (1984) *Nucl. Acids Res.* 12: 387-395), DNA Strider 1.2 (March (1991) DNA Strider 1.2, Gif-sur-Yvette, France), the BLAST search tool (Altschul et al. (1990) *J. Mol. Biol.* 215: 403-410) and the tRNAscan-SE 1.1 program (Lowe and Eddy (1997) *Nucl. Acids Res.* 25: 955-964).
15 Phylogenetic relationships were determined by the neighbor-joining distance algorithm using Phylip (Felsenstein (1993) PHYLIP (Phylogeny Inference Package). Department of Genetics, University of Washington, Seattle; SAITOU and NEI (1987) *Mol. Biol. Evol.* 4: 406-425) and were based on reverse transcriptase amino acid sequences that had been aligned with ClustalW1.7.
20 THOMPSON, et al. (1994) *Nucl. Acids Res.* 22: 4673-4680. Transmembrane helices were identified using the PHDhtm program. ROST et al. (1995) *Prot. Science* 4: 521-533. All DNA sequences have been submitted to the DDBS/EMBL/GenBank databases under the accession numbers X12345, X23456, X34567 and X45678.

25

RESULTS

30 Tat1 is a retrotransposon: Tat1 insertions share features with retrotransposon solo LTRs. We reasoned that if Tat1 is a retrotransposon, then there should be full-length elements in the genome consisting of two Tat1 sequences flanking an internal retrotransposon coding region. To test this hypothesis, additional Tat1 elements were isolated by screening a Landsberg (La-0) genomic DNA library with a Tat1 probe. Twenty-one λ phage clones were isolated and Southern analysis revealed two clones (pDW42 and pDW99) each with two copies of Tat1 (data not shown). The two Tat1 elements in each clone were sequenced, along with the intervening DNA. All Tat1 sequences shared >89% nucleotide identity to the previously characterized Tat1a - Tat1c elements. Peleman
35

et al. (1991) Proc. Natl. Acad. Sci. USA 88: 3618-3622. In clone pDW99, the 5' and 3' Tat1 sequences were 433 bases in length and only differed at two base positions. These Tat1 sequences also had conserved features of LTRs, including the dinucleotide end-sequences (5' TG-CA 3') that were part of 12 base inverted terminal repeats. If the two Tat1 elements in clone pDW99 were retrotransposon LTRs, then both, along with the intervening DNA, should be flanked by a target site duplication. A putative five base target site duplication (TATGT) was present immediately adjacent to the 5' and 3' Tat1 elements, supporting the hypothesis that they and the intervening DNA inserted as a single unit. In clone pDW42, the 5' Tat1 was 432 bases in length and shared 98% nucleotide sequence identity to the 3' Tat1. The last ~74 bases of the 3' Tat1 was truncated during library construction and lies adjacent to one phage arm. A target site duplication, therefore, could not be identified in this clone.

DNA sequences were analyzed for potential coding information between the 5' and 3' Tat1 elements. Nearly identical ORFs of 424 and 405 amino acids were found encoded between the Tat1 sequences in pDW42 and pDW99, respectively. The derived amino acid sequences of these ORFs were used to search the DNA sequence database with the BLAST search tool, and significant similarity was found to the Zea mays retrotransposable element Zeon-1 ($p = 4.4e-08$). HU et al. (1995) Mol. Gen. Genet. 248: 471-480. The ORFs have ~44% similarity across their entirety to the 628 amino acid ORF encoded by Zeon-1 (see below). The Zeon-1 ORF includes a zinc finger motif characteristic of retrotransposon gag protein RNA binding domains. Hu et al. (1995) Mol. Gen. Genet. 248: 471-480. Although the Tat1 ORFs do not include the zinc finger motif, the degree of similarity suggests that they are part of a related gag protein.

If the Tat1 sequences in pDW42 and pDW99 defined retrotransposon insertions, a PBS would be predicted to lie adjacent to the 5' Tat1 elements in both clones. The putative Tat1 PBS shares similarity with PBSs of Zeon-1 and another maize retrotransposon called Cinful (see below), but it is not complementary to an initiator methionine tRNA as is the case for most plant retrotransposons. Additionally, a possible polypurine tract (PPT), the primer for second strand cDNA synthesis, was observed one base upstream of the 3' Tat1 sequence in both phage clones (5'-GAGGACTTGGGGGGCAAA-3'). We concluded from the available evidence that Tat1 is a retrotransposon, and we have designated the 3960 base

insertion in pDW42 as Tat1-1 and the 3879 base insertion in pDW99 as Tat1-2. It is apparent that both Tat1-1 and Tat1-2 are non-functional. Their ORFs are truncated with respect to the coding information found in transposition-competent retrotransposons, and they lack obvious pol motifs.

5

In light of our findings, the previously reported Tat1 sequences can be reinterpreted. Tat1a and Tat1b, which are flanked by putative target site 10 duplications, are solo LTRs. Tat1c, the only element without a target site duplication, is actually the 5' LTR and part of the coding sequence for a larger Tat1 element.

10

Copy number of Tat1 among *A. thaliana* ecotypes: To estimate Tat1 copy 15 number, the 5' LTR, gag and the 3' non-coding region were used as separate probes in Southern hybridizations. The Southern filters contained genomic DNA from 17 ecotypes representing wild populations of *A. thaliana* from around the world. This collection of ecotypes had previously been used to evaluate 20 retrotransposon population dynamics. Konieczny et al. (1991) Genetics 127: 801-809; Voytas et al. (1990) Genetics 126: 713-721; Wright et al. (1996) Genetics 142: 569-578. Based on the hybridization with the gag probe, element copy 25 number ranges from two to approximately ten copies per ecotype. The copy number of the LTRs is higher, likely due to the presence of two LTRs flanking full-length elements or solo LTRs scattered throughout the genome. The Tat1 copy number contrasts with the copy numbers (typically less than three per ecotype) observed for 28 other *A. thaliana* retrotransposon families. Konieczny et al. (1991) Genetics 127: 801-809; Voytas et al. (1990) Genetics 126: 713-721; Wright et al. (1996) Genetics 142: 569-578. In addition, the Tat1-hybridizing restriction fragments are highly polymorphic among strains. This degree of polymorphism, coupled with the high copy number, suggested that Tat1 has been active in transposition since the separation of the ecotypes.

30

The Tat1 3' non-coding region contains DNA sequences from elsewhere in 35 the genome: In an attempt to identify a complete and functional Tat1 element, LTR-specific primers were used in PCR reactions optimized for amplification of large DNA fragments. Most full-length retrotransposable elements are between five and six kb in length. DNAs from all 17 ecotypes were used as templates, and each gave amplification products of ~3.2 kb, the size predicted for Tat1-1 and Tat1-2 (data not

shown). In La-0, however, a 3.8 kb PCR product was also recovered. This PCR product was cloned, sequenced and called Tat1-3. This insertion is expected to be about 4.6 kb in total length if the LTR sequences are included.

5 Tat1-3 differed from Tat1-1 and Tat1-2 in that it had two ORFs separated by stop codons and a 477 base insertion in the 3' non-coding region. The first ORF (365 amino acids) was similar to but shorter than the ORFs of the other Tat1 elements. The sequences constituting the second ORF (188 amino acids) were not present in the other Tat1 insertions and were not related to other sequences in the 10 DNA databases. Database searches with the 477 base insertion in the 3' non-coding region, however, revealed three regions of similarity to other genomic sequences. A region of 113 bases matched a region of 26 bp repeats in the 5' untranslated sequence of the AT-P5C1 mRNA, which encodes pyrroline-5-carboxylate reductase ($p = 2.1e-19$). Verbruggen et al. (1993) *Plant Physiol.* 103: 15 771-781. In addition, 50 bases appear to be a remnant of another retrotransposon related to Tat1. These 50 bases are 71% identical to the 3' end of the Tat1-3 LTR and the putative primer binding site. The putative primer binding site, however, is more closely related to those of other plant retrotransposons such as Huck-2 (Sanmiguel et al. (1996) *Science* 274: 765-768). Finally, sequences in the 20 remainder of the insertion showed significant similarity to a region on chromosome 5. To confirm that Tat1-3 was not a PCR artifact, two additional primer pairs were used in separate amplifications. Both amplifications gave PCR products of the predicted sizes, which were cloned and confirmed to be Tat1-3 by DNA sequencing.

25 PCR amplifications with the additional primer pairs also yielded a product 0.8 kb longer than that expected for Tat1-3. This product was cloned, sequenced and found to be another Tat1 element, designated Tat1-4. This element has sequences similar to a Tat1 LTR, polypurine tract and the second ORF of Tat1-3. 30 In Tat1-4, 1182 bases of DNA are found in the 3' non-coding region at the position corresponding to the 477 base insertion in Tat1-3. This region does not match any sequences in the DNA databases.

35 Other Tat1-like elements in *A. thaliana*: A BLAST search of DNA sequences generated by the *A. thaliana* genome project identified two more solo LTRs similar to Tat1. All share similarities throughout, but most strikingly, they

are very well conserved at the 5' and 3' ends where it is expected integrase would bind. Braiterman and Boeke (1994) Mol. Cell. Biol. 14: 5731-5740. These conserved end-sequences suggest that the integrases encoded by full-length elements are also related, and that the LTRs have evolved under functional constraints; that is, they are not simply degenerate Tat1 LTRs. The two new LTRs are designated as Tat2-1 and Tat3-1. Tat2-1 is 418 bases long, is flanked by a five base target site duplication (CTATT) and is ~63% identical to the Tat1-2 5' LTR. Tat3-1 is 463 bases long and is also flanked by a target site duplication (ATATT). Tat3-1 is ~53% identical to the Tat1-2 5' LTR.

5

Tat1 and Athila are related to Ty3/gypsy retrotransposons: Further analysis of data from the *A. thaliana* genome project revealed two slightly degenerate retrotransposons with similarity to the Tat1 ORF. These elements were identified within the sequence of the P1 phage clones MXA21 (Accession AB005247; bases 10
15
54,977-66,874) and MX110 (Accession AB005248; bases 24,125-35,848). Each has two LTRs, a putative PBS, and long ORFs between their LTRs. The genetic organization of these elements is depicted in Figures 5A and 6A. Amino acid sequence analysis indicated the presence of an RNA binding domain that defines gag in both elements. This region is followed by conserved reverse transcriptase, RNaseH, and integrase amino acid sequence domains characteristic of pol (data not shown). Classification of eukaryotic retrotransposons into the Ty1/copia elements (Pseudoviridae) and Ty3/gypsy elements (Metaviridae) is based on pol gene structure. Boeke et al. (1998) Metaviridae. In Virus Taxonomy: ICTV VIIth Report, edited by F. A. Murphy. Springer-Verlag, New York.; Boeke et al. (1998b) Pseudoviridae. In Virus Taxonomy: ICTV VIIth Report, edited by F. A. Murphy. Springer Verlag, New York. The domain order of the pol genes (reverse transcriptase precedes integrase) and similarities among their encoded reverse transcriptases (see below) identifies these elements as the first full-length *A. thaliana* Ty3/gypsy elements.

20
25
30

Because the characterized Tat1 insertions do not encode pol genes, this element family could not be classified. However, the amino acid sequence of the Tat1-2 ORF is 51% similar to the gag region of the MXA21 retrotransposon. Since plant retrotransposons within the Ty1/copia or Ty3/gypsy families, even those with highly similar pol genes, share little amino acid sequence similarity in their gag regions, Tat1 is likely a Ty3/gypsy element. This conclusion is further supported

5 by the report that the Tat-like Zeon-1 retrotransposon is very similar to a *Z. mays* Ty3/gypsy element called cinful (Bennetzen (1996) Trends Microbiol. 4: 347-353); however, only the 5' LTR and putative primer binding site (PBS) sequences are available in the sequence database for analysis (Accession U68402). Because of the extent of similarity to Tat1, we have named the MXA21 insertion Tat4-1.

10 The gag region of the MX110 element is 62% similar ($p = 1.1e-193$) to the first ORF of Athila, which has previously been unclassified (Pelissier et al. (1995) Plant Mol. Biol. 29: 441-452). This implies that Athila is also a Ty3/gypsy element, and we have designated the MX110 insertion as Athila1-1. Our classification of Athila as a Ty3/gypsy element is further supported by the observation that the Athila gag amino acid sequences shares significant similarity to the gag protein encoded by the cyclops-2 Ty3/gypsy retrotransposon of pea (Accession AJ000640; $p = 1.1e-46$; data not shown). Further analysis of the 15 available *A. thaliana* genome sequences identified three additional Athila homologs. They include an additional Athila1 element, designated Athila1-2, and two more distantly related Athila-like elements, designated Athila2-1 and Athila3-1.

20 In addition to similarities among their gag amino acid sequences, the Tat elements have short LTRs (<550 bp) and long 3' non-coding regions (>2 kb). In contrast, the Athila-like elements have long LTRs (>1.2 kb) and are very large 25 retrotransposons (>11 kb). One additional feature to note about both the Athila-like and Tat-like elements is the high degree of sequence degeneracy of their internal coding regions. This contrasts with the near sequence identity of their 5' and 3' LTRs, which is typically greater than 95%. Because a single template is used in the 30 synthesis of both LTRs, LTR sequences are usually identical at the time of integration. The degree of sequence similarity between the LTRs suggests that most elements integrated relatively recently. The polymorphisms observed in the internal domains of these insertions, therefore, may have been present in their progenitors, and these elements may have been replicated in trans.

35 A novel, conserved coding region in Athila elements: A surprising feature of Athila1-1 is the presence of an additional ORF after integrase. Like gag, this ORF shares significant similarity across its entirety ($p = 3.8e-08$) to the second ORF of Athila. This ORF is also encoded by the Athila2-1 and Athila3-1 elements, although it is somewhat more degenerate. The presence of this coding sequence

among these divergent retrotransposons suggests that it plays a functional role in the element replication cycle. However, the ORF shows no similarity to retrotransposon gag or pol genes. The retroviruses and some Ty3/gypsy retrotransposons encode an env gene after integrase. Although not well-conserved in primary sequence, both viral and retrotransposon envelope proteins share some structural similarities. They are typically translated from spliced mRNAs and the primary translation product encodes a signal peptide and a transmembrane domain near the C-terminus. All four families of Athila elements encode a domain near the center of the ORF that is strongly predicted to be a transmembrane region (70% - 90% confidence, depending on the element analyzed) (ROST et al. (1995) *Prot. Science* 4: 521-533). Two retrotransposons, Athila and Athila2-1, also have a hydrophobic transmembrane domain near the 5' end of their env-like ORFs, which may serve as a secretory signal sequence. Von Heijne (1986) *Nucl. Acids Res.* 14: 4683-4690.

Two lineages of plant Ty3/gypsy retrotransposons: Relationships among Ty3/gypsy retrotransposons from *A. thaliana* and other organisms were assessed by constructing a neighbor-joining tree of their reverse transcriptase amino acid sequences. Included in the analysis were reverse transcriptases from two additional families of *A. thaliana* Ty3/gypsy elements that we identified from the unannotated genome sequence data (designated Tma elements; Tma1-1 and Tma3-1); two other Tma element families were identified in the genome sequence that did not encode complete reverse transcriptases (Tma2-1 and Tma4-1; Table 1). Also included in the phylogenetic analyses were reverse transcriptases from a faba bean retrotransposon and the cyclops-2 element from pea. The plant Ty3/gypsy group retrotransposons resolved into two lineages: One was made up of dell1 from lily, the IFG7 retrotransposon from pine, reina from *Z. mays*, and Tma1-1 and Tma3-1. This group of elements formed a single branch closely related to numerous fungal retrotransposons (branch 1). The second branch (branch 2) was well-separated from all other known Ty3/gypsy group elements, and was further resolved into two lineages: Athila1-1, cyclops-2 and the faba bean reverse transcriptase formed one lineage (the Athila branch), and Tat4-1 and Grande1-4 from *Zea diploperennis* formed a separate, distinct branch (the Tat branch).

Primer binding sites: Most plant Ty1/copia retrotransposons as well as the branch 1 Ty3/gypsy elements have PBSs complementary to the 3'-end of an

5 initiator methionine tRNA. This is not the case for any of the branch 2 Ty3/gypsy elements. We compared the putative PBSs of Tat-branch and Athila-branch elements to known plant tRNA genes as well as to the 11 tRNA genes that had been identified to date in sequences generated by the *A. thaliana* genome project. In addition, we searched the unannotated *A. thaliana* genome sequences and identified 30 more *A. thaliana* tRNA genes using the program tRNAscan-SE (Lowe and Eddy 10 (1997) *Nucl. Acids Res.* 25: 955-964). The PBS of Tat1 is complementary to 10 bases at the 3' end of the asparagine tRNA for the AAC codon; these 10 bases are followed by a two base mismatch and six additional bases of perfect complementarity. The Tat4-1 PBS is complementary to 20 bases at the 3' end of the arginine tRNA for the AGG codon with one mismatch 10 bases from the 3' end; Huck-2, Grande-zm1, Grande1-4, and the retrotransposon-like insertion in the 3' 15 non-coding region of Tat1-3 all have 20-base perfect complementarity to this tRNA. The PBS of Athila1-1 is perfectly complementary to 15 bases at the 3' end of the aspartic acid tRNA for the GAC codon, and Athila and Athila2-1 have 13 bases of complementarity to this tRNA. At this time there is no known plant tRNA complementary to the PBS of Zeon-1, which has the same PBS as the maize retrotransposon *ciful*. As more tRNA sequences become available, a candidate primer may be identified for these elements.

20

Example 2 Characterizing the *Pisum sativum* Retroelement ("Cyclops" element) env gene

25 After identifying the retrovirus-like elements in *A. thaliana*, the element called Cyclops2 from *Pisum sativum* (Chavanne et al. (1998) *Plant Mol. Biol.* 37:363-375) was examined. Comparison of this element to the Athila-like elements both in size and amino acid and nucleotide sequence composition was made. Cyclops2 also encodes an open reading frame (ORF) in the position corresponding to the env-like gene of the Athila elements. This Cyclops2 ORF was examined 30 using the same methods used to characterize the Athila group env-like genes (see Example 1). The Cyclops2 ORF was found to have a potential splice site at its N-terminus and transmembrane domains at the N-terminus, the central region and the C-terminus. Based on the presence of these features, it was concluded that Cyclops2 is a retrovirus-like retroelement that encodes an env-like gene.

Example 3 Obtaining the Soybean Retroelements (“Calypso” elements)

Materials and Methods

Library Screening and Southern Hybridization. A soybean genomic lambda phage library (line L85-3044) was initially screened with a reverse transcriptase probe under low stringency conditions (50 degrees Celsius with a 1% SDS wash) (Church and Gilbert (1984) Proc. Natl. Acad. Sci. USA 81:1991-1995). The library was previously described (Chen et al. (1998) Soybean Genetics Newsletter 25:132-134). The probe was obtained by PCR amplification of genomic *P. sativum* DNA using primers based on the reverse transcriptase of Cyclops2 (DVO701 and DVO702). All probes were radio-labeled using random primers and protocols supplied by Promega (Madison, WI). For Southern hybridizations, DNA was digested, electrophoresed on 0.8% agarose gels, and transferred to Gene Screen Plus membranes using the manufacturer's alkaline transfer protocol (New England Nuclear, Boston, MA). All high stringency hybridizations were as described (Church and Gilbert (1984) Proc. Natl. Acad. Sci. USA 81:1991-1995).

DNA sequencing. Lambda phage clones were subcloned into the vector pBluescript KSII - and transformed into the *E.coli* host strain XL1 Blue (Stratagene, La Jolla, CA) (Ausubel et al., Current Protocols in Molecular Biology (Greene Publishing Associates, Inc., 1993). Subclones were sequenced by primer walking at the Iowa State University DNA sequencing facility.

Sequence Analysis. DNA Sequence analysis was performed using the GCG software package (Devereux et al. (1984) Nucleic Acids Res. 12:387-395), DNA Strider 1.2 (Marck (1991) DNA Strider 1.2, Gif-sur-Yvette, France) and the BLAST search tool (Altschul et al. (1990) J. Mol. Biol. 215: 403-410). Phylogenetic relationships were determined by the neighbor-joining distance algorithm (Saitou and Nei (1987) Mol. Biol. Evol. 4: 406-425) using PAUP v4.0 beta 1 (Swofford (1993) Illinois Natural History Survey, Champaign, IL) and were based on reverse transcriptase amino acid sequences that had been aligned with ClustalX v1.63b (Thompson et al. (1994) Nucl. Acids Res. 22: 4673-4680). Transmembrane helices were identified using the PHDhtm program and TMpred

(Rost et al. (1995) *Prot. Science* 4: 521-533; Hofmann and Stoffel (1993) *Biol. Chem.* 374:166).

Results

Retrovirus-like elements in *Glycine max*. Soybean retrovirus-like elements were identified by a low stringency (50 degrees C) screen of a soybean lambda library using a reverse transcriptase probe. The probe was based on a sequence from Cyclops2 (Chavanne et al. (1998) *Plant Mol. Biol.* 37:363-375). The screen produced 63 lambda clones that appeared to contain a retrovirus-like reverse transcriptase based on hybridization to the probe. Thirty-five of these putative elements were sequenced to varying degrees and 24 encoded readily identifiable retrovirus-like sequences. Most of the elements were distantly related and had premature stop codons, frame shifts, deletions or insertions. A related group of three elements and another related pair were completely sequenced and analyzed. The three elements in the first group are referred to as Calypso1-1, Calypso1-2, and Calypso1-3. The elements in the second pair are referred to as Calypso2-1 and Calypso2-2. The remaining soybean retrovirus-like elements will be given the Calypso name and a sequential designator number based on their family grouping.

The Calypso retrovirus-like elements have the same overall structure and sequence homology as the previously described Athila and Cyclops elements. The elements are ~12kb in length; they have a 5' LTR, a PBS (Primer Binding Site), a gag protein, a pol protein, a spacer, an env-like protein, another spacer region, a PPT (Polypurine Tract) and a 3' LTR. The LTRs vary from ~1.3 to ~1.5kb in length and characteristically begin with TG and end with CA. The PBS is similar to that used by the Athila and Cyclops elements; it is 4 to 6 bases past the 5' LTR and matches the 3' end of a soybean aspartic acid tRNA for 18 to 19 bases with 1 mismatch. The fact that the sequences of the Calypso primer binding sites are shared with the *A. thaliana* and *P. sativum* retrovirus-like elements, indicates that this sequence is a unique marker for envelope-encoding retroelements. The gag protein extends ~850 amino acids and encodes a zinc finger domain (characterized by the amino acid motif CxxCxxxHxxxxC) and a protease domain (characterized by the amino acid motif LIDLGA). These domains are located at approximately the same positions within gag as in other retroelements. The ~600 amino acid reverse transcriptase region follows gag and has the conserved plant retrovirus-like motifs which approximate the following amino acids: KTAF, MP/SFGLCNA,

V/I/MEVFMDDFS/WV/I, FELMCDASD YAI/VGAVLGQR, and YATT/IEKEL/MLAIVF/YAL/FEKFR/KSYLI/VGSR/KV, respectively. The ~450 amino acid integrase domain has the plant retrovirus-like integrase motifs that approximate HCHxSxxGGH30xCdxCQR for the Zn finger as well as two other motifs that approximate WGIDFI/V/MGP, and PYHPQTxGQA/VE. After integrase, there is a ~0.7kb spacer then a ~450 amino acid env-like protein coding region. The env-like protein of the Calypso elements is well conserved through most of the ORF but conservation decreases toward the C-terminus. The conservation includes 2 or 3 presumed transmembrane domains and a putative RNA splice site acceptor. The env-like protein is followed by a ~2 kb spacer then a polypurine tract with the approximate sequence ATTTGGGGG/AANNT. The 3' LTR starts immediately after the final T of the PPT.

Calypso elements are abundant and heterogeneous. The Calypso elements appear to be abundant in the soybean genome. High stringency Southern blots of soybean DNA probed with reverse transcriptase, gag or env-like sequences produced smeared hybridization patterns, suggesting that the elements are abundant and heterogeneous. Their heterogeneity was also supported by DNA sequence analysis, which revealed a maximum of 93% nucleotide identity among elements, and most elements averaged ~88% nucleotide identify. This identity can be region-specific or dispersed over the element's entirety. For example, reverse transcriptase, integrase and envelope-like coding regions may be well conserved, whereas the LTR, gag and spacer regions may have very little sequence conservation.

Phylogenetic analysis of Calypso reverse transcriptase. The reverse transcriptase of retroelements is the preferred protein for assessment of phylogenetic relationships (Xiong and Eickbush (1990) EMBO J. 9:3353-3362). This is due to the high degree of amino acid sequence conservation found in reverse transcriptase proteins from many sources. The Calypso retrovirus-like elements were compared to previously described Ty3/gypsy and retrovirus-like elements from plants, fungi and invertebrate animals. The Calypso elements formed a distinct group with other plant retrovirus-like elements from *A. thaliana* and *P. sativum* and Faba bean. This group did not include plant Ty3/gypsy elements that are members of the metaviridae genus. This indicates that the plant retrovirus-like

elements from these four plant species are closely related and form a new element group that may be present in all or most plant species.

The Calypso reverse transcriptase and integrase are well-conserved. Frame shifts in the retrovirus-like elements were repaired through sequence comparison between the retrovirus-like elements from *A. thaliana*, *P. sativum* and *G. max*. Restoration typically involved an insertion or deletion of a single nucleotide or a single nucleotide substitution. When the edited ORFs of seven plant retrovirus-like elements from three species were compared, it was found that the gag domain had very little conservation. The amino acid sequence around the protease domain was reasonably conserved (~50%) but the reverse transcriptase and integrase domains were highly conserved (~70%).

The env-like ORF of Calypso is well-conserved. Animal retrovirus env proteins share little in common. They are however cleaved into two functional units that consist of the surface (SU) and transmembrane (TM) peptides. The SU peptide contains a transmembrane secretory signal at the N-terminus. The TM peptide has two transmembrane domains, one at the N-terminus, which functions in membrane fusion, and another near the C-terminus, which acts as an anchor site. The retrovirus env protein is expressed from an RNA that is spliced near the beginning of the env ORF. There are currently nine Athila group elements from *A. thaliana* that have an identifiable env-like ORF. Alignment of the env-like amino acid sequence shows that there are five subgroups of env-like proteins in the Athila family. Three are distinct, four are closely related and another pair is closely related. As a whole, these env-like sequences share limited homology over the entire length of the ORF, but within subgroups, they share high homology (data not shown). Some of the Athila env-like proteins have an apparent secretory peptide and a central transmembrane domain, suggesting that they may have an env-like function.

Among the Calypso elements, seven have been characterized that encode env-like ORFs. These env-like ORFs form four families that have a high degree of overall sequence similarity beginning at the first methionine and continuing for three quarters of the ORF; sequence similarity falls off dramatically near the C-terminus. The amino acid sequence at the first methionine has the consensus sequence QMASR/KKRR/KA, which appears to be a nuclear targeting signal, however, the

program PSORT only predicts a 0.300 confidence level for this targeting role (Nakai and Horton (1999) Trends Biochem. Sci. 24:34-36). A similar sequence (ASKKRK) is found at the same position in the env-like ORF of Cyclops2, suggesting that it serves a similar purpose. No other potential targeting peptide stands out from the sequence that has been analyzed so far. There is a conserved region that is predicted to be a transmembrane domain near the center of the Calypso env-like protein and a second transmembrane domain located at variable positions near the C-terminus. These may be the fusion and anchor functions of a TM peptide. It should also be noted that five of the seven ORFs are predicted to have a transmembrane domain that is just before and includes the first methionine. This N-terminal transmembrane domain may be a secretory signal of an SU peptide. The program TMpred estimates these transmembrane domains to be significant based on a score >500 (Hofmann and Stoffel (1993) Biol. Chem. 374:166). These three transmembrane domains are found in the Cyclops2 env-like protein at similar locations but at a reduced significance score. Another feature of the Calypso env-like ORF is the conserved splice site that is predicted to be at the first methionine by the program NetGene2 v. 2.4 with a confidence level of 1.00 (Hebsgaard et al. (1996) Nucleic Acids Res. 24:3439-3452); Brunak et al. (1991) J. Mol. Biol. 220:49-65). There are other less preferred putative splice sites in the region, but only the splice site near the methionine is optimally placed and conserved in all seven env-like ORFs.

Example 4 Obtaining the Generic Plant Retroelements ("Generic" elements)

ClustalX v1.63b (Thompson et al. (1994) Nucl. Acids Res. 22: 4673-4680) was used to align nucleotide sequences of Calypso1-1, Calypso1-2 and Calypso1-3. A consensus sequence was generated from the ClustalX output. The consensus sequence file was then translated and compared using ClustalX to amino acid sequences of retrovirus-like elements from soybean, pea (Cyclops2) and *A. thaliana* (Athila-like elements) using the GCG computer software package (Devereux et al. (1984) Nucleic Acids Res. 12:387-395). For coding regions encompassing protease, reverse transcriptase and integrase, a new consensus sequence was generated that best matched the coding information in all elements. This second consensus sequence forms the protease, reverse transcriptase and integrase genes of the generic element. The gag gene of the generic element is a consensus sequence

generated by editing alignments between Calypso1-1 and Calypso2-2. The env gene is a consensus sequence based on env gene sequence alignments of all Calypso elements. All non-coding regions for the generic element were obtained >from Calypso1-2, with the exception of the LTRs, which were taken from Calypso1-1.

5

A generic retrovirus will be constructed by first generating a DNA sequence that approximates the sequence of the generic element. An element that closely matches the consensus -- for example, Calypso1-1 -- will be modified by PCR-based site-directed mutagenesis (Ausubel et al., Current Protocols in Molecular Biology (Greene Publishing Associates, Inc., 1993). Modifications will be sequentially introduced into the starting element until it conforms to the sequence of the generic element.

10

The generic element will be modified so that it will be expressed at high levels in plant cells. This will be accomplished by inserting an enhancer -- such as the cauliflower mosaic virus 35S enhancer -- into the 5' LTR. To monitor replication, a marker gene will be inserted into the virus between the end of the coding region for the env gene and the polypurine tract. The marker gene may encode resistance to an herbicide or antibiotic. The modified generic element will then be introduced into plant cells by standard means of plant transformation. Because the modified generic element will be expressed at high levels, retroviral particles will be produced by the host plant cell. These will be harvested and purified by passing cell lysates over sucrose density gradients.

15

The plant retroviral particles will be incubated in the presence of non-transformed plant cells. The virus will associate with the plant cell and fuse with the plant cell membrane. The mRNA carried by the virus will be reverse transcribed and the resultant cDNA will be integrated into the genome of the plant. The integration of the viral DNA and the expression of the marker gene it carries will confer antibiotic resistance to the plant cell. Cells that carry integrated viruses can be identified through genetic selection.

20

30

35

Although the present invention has been fully described herein, it is to be noted that various changes and modifications are apparent to those skilled in the art. Such changes and modifications are to be understood as included within the scope of the present invention as defined by the appended claims.